

Comprehensive Clinical Psychology. Volume 4

Copyright © 2000 Elsevier Science Ltd. All rights reserved.

Editors-in-Chief: Alan S. Bellack and Michel Hersen

Table of Contents

Volume 4: Assessment

 Preface

 Contributors

4.01 The Role of Assessment in Clinical Psychology, Pages 1-32, Lee Sechrest, Timothy R. Stickle and Michelle Stewart
SummaryPlus | Chapter | PDF (422 K)

4.02 Fundamentals of Measurement and Assessment in Psychology, Pages 33-55, Cecil R. Reynolds
SummaryPlus | Chapter | PDF (372 K)

4.03 Diagnostic Models and Systems, Pages 57-80, Roger K. Blashfield
SummaryPlus | Chapter | PDF (371 K)

4.04 Clinical Interviewing, Pages 81-96, Edward L. Coyle Diane J. Willis, William R. Leber and Jan L. Culbertson
SummaryPlus | Chapter | PDF (319 K)

4.05 Structured Diagnostic Interview Schedules, Pages 97-130, Jack J. Blanchard and Seth B. Brown
SummaryPlus | Chapter | PDF (437 K)

4.06 Principles and Practices of Behavioral Assessment with Children, Pages 131-155, Thomas H. Ollendick Ross W. Greene
SummaryPlus | Chapter | PDF (366 K)

4.07 Principles and Practices of Behavioral Assessment with Adults, Pages 157-186, Stephen N. Haynes
SummaryPlus | Chapter | PDF (414 K)

4.08 Intellectual Assessment, Pages 187-238, Alan S. Kaufman Elizabeth O. Lichtenberger
SummaryPlus | Chapter | PDF (588 K)

4.09 Assessment of Memory, Learning, and Special Aptitudes, Pages 239-265, Robyn S. Hess Rik Carl D'amato
SummaryPlus | Chapter | PDF (366 K)

4.10 Neuropsychological Assessment of Children, Pages 267-301, Cynthia A. Riccio and Cecil R. Reynolds

SummaryPlus | Chapter | PDF (420 K)

4.11 Neuropsychological Assessment of Adults, Pages 303-347, C. Munro Cullum

SummaryPlus | Chapter | PDF (786 K)

4.12 Principles of Personality Assessment, Pages 349-370, Jerry S. Wiggins and Krista K. Trobst

SummaryPlus | Chapter | PDF (381 K)

4.13 Observations of Parents, Teachers, and Children: Contributions to the Objective Multidimensional Assessment of Youth, Pages 371-401, David Lachar

SummaryPlus | Chapter | PDF (453 K)

4.14 Objective Personality Assessment with Adults, Pages 403-429, James N. Butcher and Jeanette Taylor G. Cynthia Fekken

SummaryPlus | Chapter | PDF (403 K)

4.15 Projective Assessment of Children and Adolescents, Pages 431-458, Irving B. Weiner and Kathryn Kuehnle

SummaryPlus | Chapter | PDF (405 K)

4.16 Assessment of Schema and Problem-solving Strategies with Projective Techniques, Pages 459-499, Hedwig Teglassi

SummaryPlus | Chapter | PDF (479 K)

4.17 Computer Assisted Psychological Assessment, Pages 501-523, Gale H. Roid and W. Brad Johnson

SummaryPlus | Chapter | PDF (347 K)

4.18 Therapeutic Assessment: Linking Assessment and Treatment, Pages 525-561, Mark E. Maruish

SummaryPlus | Chapter | PDF (467 K)

4.19 Forensic Assessment, Pages 563-599, David Faust

SummaryPlus | Chapter | PDF (457 K)

Preface Volume 4

Psychology is often described or defined as the science of human behavior. Science is a process of systematic, planned study and investigation. The process of science requires the ability to measure, observe, and classify phenomena of interest. The basic psychological sciences that underlie clinical

practice in psychology rely routinely on the ability to measure and assess whatever variables are of interest. As our ability to measure more variables and to do so more accurately has progressed, so has science and practice in psychology.

The beginnings of psychology as a science are commonly attributed to the experimental laboratory of Wilhelm Wundt in Leipzig, where work was based largely on the measurement of sensory processes. One of the key reasons Wundt is credited with founding scientific psychology is because of his emphasis on objective measurement. Lightner Witmer, who must have been the leading "Renaissance scholar" in the psychology of his day, is credited by various chroniclers of the discipline as the founding father of clinical psychology, school psychology, and clinical neuropsychology. Witmer was strongly influenced by Wundt and his approach of objective measurement and analysis and the instruction he received from another experimental psychologist, E. B. Twitmeyer (whose discovery of classical conditioning predated that of Pavlov). In his early works, Witmer describes the process of mental analysis as one founded in the experimental nature of science (Witmer, 1902) tempered with the knowledge of human development and careful observation in a manner surprisingly coincident with the modern-day approach of Kaufman (1994). Witmer subsequently founded the first recorded psychological clinic for children, at the University of Pennsylvania, and began an experimental school for children with disabilities, known then as backwards children. Witmer remarked often about the need to integrate knowledgeable observation with careful measurement to produce an assessment of the child that leads to insights about interventions. This remains our goal, even though our methods are more sophisticated. It was through his work at The Psychological Clinic that Witmer germinated so much of what is professional psychology today.

Clinical psychology, school psychology, and clinical neuropsychology can all trace their roots to the unique psychological skills of reproducible assessments of human behavior. The school systems began to need to classify pupils for differentiated instruction and pioneers such as Dorothea Dix altered public policy toward the mentally ill, creating a need for more accurate differential diagnosis. Simultaneously, the military of the United States needed to assess and classify thousands of records, find those mentally unfit for duty, and treat the mental casualties of service. All of these activities required the unique skills of the clinician in diagnostic psychological testing. Our medical colleagues gradually began to recognize the value of psychological testing for differential diagnosis of mental disorders. As our diagnostic skills have progressed, so the diagnosis and classification of mental disorders through formal taxonomies (e.g., the *International classification of diseases and the Diagnostic and statistical manual of mental disorders*) has become more objective.

Our ability to engage in actuarial diagnosis and decision-making has increased geometrically with the inexpensive availability of personal computers. This technology is ahead of practice as is usually the case, yet one cannot help but observe that psychology is perhaps slower than most clinical professions to adopt such changes. Perhaps it is due to our charge to care for the human psyche that causes us to hold on to more personalized approaches to diagnosis. Nevertheless, Witmer's prompt to use objective measurement as the foundation of clinical practice seems forever sound, and it is to this end this volume is intended.

This volume of *Comprehensive Clinical Psychology* is devoted to an explication of the models and methods of assessment in clinical psychology, and to the varied roles the clinician encounters. From the singular office practice to the medical clinic to the courtroom, objective measurement and

assessment seems always to improve what we do. Yet, as we learn in the opening chapters, perhaps we do not take appropriate advantage of what we know and how to do it. The models and methods for doing so are expounded in the chapters that follow. Controversial approaches are analyzed and discussed (e.g., projective assessment, intelligence testing), just as are the more currently acceptable models of behavioral assessment. Links to the future and to treatment are noted throughout the volume. In all cases, science first, the empirical basis of practice is emphasized.

The volume is organized and authors chosen to produce a work in line with these philosophies. The opening chapter by Sechrest, a measurement and a personality scientist, Stickle, and Stewart, acts as gadfly to the work with their candid view of the way assessment is used in practice. This is followed by a review of the fundamental psychometrics that underlie clinical assessment, emphasizing the link between science and practice. Next, Blashfield reviews the state and evolution of taxonomies in clinical psychology and their use in the field.

A superb clinician and noted researcher, Willis and colleagues were chosen to review the role and method of the interview in clinical psychological assessment, always presenting approaches with sound backing in the literature of the discipline. Interviewing is an assessment technique, one from which we draw inferences about patients, and the validity of interview-based inferences should always be of concern. Therefore, structured interview schedules are next reviewed by Blanchard and Brown. Structured interview schedules are more easily evaluated empirically, since they often yield directly quantified results. This quantitatively oriented approach to the interview leads well to the next two chapters on behavioral assessment by Ollendick and Greene (children) and Haynes (adults). Both Ollendick and Greene have a long history of empirical research and, in their own roles as journal editors, are particularly sensitive to the role of empirical validation of the interpretations made of assessment data.

Traditional cognitive approaches to assessment are next featured and again authors have been chosen to reflect the application of measurement methods to the daily problems of clinical problems. This section begins with a review of intellectual assessment by Kaufman and Lichtenberger. Kaufman pioneered the application of statistical approaches to the evaluation of Wechsler profiles, and statistical models elaborated by sound knowledge of developmental theory and of differential psychology coupled with skilled observation. The remaining authors in this section, presenting the evaluation of memory and learning (Hess and D'Amato), and the neuropsychological integrity of children (Riccio and Reynolds) and adults (Cullum), reflect a method consistent with the research-based practices of Kaufman, but each with their own blend of research and clinical skills.

The next three chapters are devoted to objective assessments of personality. In Wiggins and Trobst's, Lachar's, and Butcher, Taylor, and Fekken's chapters, the reader will recognize names long associated with empirical models of test interpretation. For the task of presenting projective assessment from the viewpoint of data and strong theoretical models, Drs. Weiner and Kuehnle (children) and Teglassi (adults) were chosen. The pull toward idiographic, anamnestic views of projective test responses is strong, yet, in these well-published authors is found a careful, reasoned approach to these controversial methods. Steeped first in theory but followed by research, Weiner and Kuehnle and then Teglassi provide two of the most literate and sound treatments of these techniques available.

Next, Roid, a measurement scientist who has worked for well-known clinical test publishing companies but also independently as an author of tests and computer interpretive programs, and Johnson provide a strong overview of the use of the computer in assisting the clinician in evaluating test performance. This area is destined to grow as the hardware advances of late make true expert systems viable on the office PC of the clinician. Maruish, known in the field for his emphasis on looking toward outcomes, reviews the linkage between assessment and treatment. Here we are also reminded of the need to document empirically that what we do works, that our patients really do get better. Finally, Faust brings us into the courtroom with a detailed analysis of the psychologist as an expert in legal proceedings. Faust has long been a critic of expert opinion from the gut and a proponent of science and sound logic in clinical work. Outside of the journal review process, there is likely no other domain wherein one's work is subjected to such scrutiny. Although Faust's focus is on using empirically supported approaches to developing opinions and practices in forensic settings, much of what he tells us is applicable to our day-to-day office practice. All of these authors were chosen in part for their knowledge and respect of the process of science but also because they know of the practical problems we face as clinicians. They echo my own philosophy to varying degrees. Where science is available, science rules practice. Where not, strong theories are preferred over clinical intuition and anecdotal knowledge bases.

In such a large work, the efforts and cooperation of many people are required. To David Hoole and Angela Greenwell at Elsevier, my special thanks for your patience and assistance in chasing both details and manuscripts. The hard work of Alan Bellack and Michel Hersen, who took their roles as Editors-in-Chief seriously and gave real feedback that improved this work, is also much appreciated. To my chapter authors go my greatest thanks however for their patience, tenacity, and willingness to accept critique, to compromise, and to revise. Thank you one and all. To my mentors, especially Alan Kaufman, Lawrence Hartlage, and Robert T. Brown, who taught me of science and of clinical skills, I will always be indebted. To my wife and friend Julia, whose compassionate care of patients in her own clinical practice will always be an inspiration, my thanks for allowing me to pursue such works as this one, for understanding the level of effort required, and for debating with me many of the ideas represented here. You make me better in all things.

References

- Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: Wiley.
 Witmer, L. (1902). *Analytical psychology*. Boston: Ginn & Company.

Volume 4 Contributors

BLANCHARD, J. J. (University of New Mexico, Albuquerque, NM, USA)
**Structured Diagnostic Interview Schedules*

BLASHFIELD, R. K. (Auburn University, AL, USA)
Diagnostic Models and Systems

BROWN, S. B. (University of New Mexico, Albuquerque, NM, USA)

**Structured Diagnostic Interview Schedules*

BUTCHER, J. N. (University of Minnesota, Minneapolis, MN, USA)

**Objective Personality Assessment with Adults*

COYLE, E. L. (Oklahoma State Department of Health, Oklahoma City, OK, USA)

**Clinical Interviewing*

CULBERTSON, J. L. (University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA)

**Clinical Interviewing*

CULLUM, C. M. (The University of Texas Southwestern Medical Center at Dallas, TX, USA)

Neuropsychological Assessment of Adults

D'AMATO, R. C. (University of Northern Colorado, Greeley, CO, USA)

**Assessment of Memory, Learning, and Special Aptitudes*

FAUST, D. (University of Rhode Island, Kingston, RI, USA)

Forensic Assessment

FEKKEN, G. C. (Queen's University, Kingston, ON, Canada)

**Objective Personality Assessment with Adults*

GREENE, R. W. (Harvard Medical School, Boston, MA, USA)

**Principles and Practices of Behavioral Assessment with Children*

HAYNES, S. N. (University of Hawaii at Manoa, Honolulu, HI, USA)

Principles and Practices of Behavioral Assessment with Adults

HESS, R. S. (University of Nebraska at Kearney, NE, USA)

**Assessment of Memory, Learning, and Special Aptitudes*

JOHNSON, W. B. (George Fox University, Newberg, OR, USA)

**Computer Assisted Psychological Assessment*

KAUFMAN, A. S. (Yale University School of Medicine, New Haven, CT, USA)

**Intellectual Assessment*

KUEHNLE, K. (University of South Florida, Tampa, FL, USA)

**Projective Assessment of Children and Adolescents*

LACHAR, D. (University of Texas-Houston Medical School, Houston, TX, USA)

Observations of Parents, Teachers, and Children: Contributions to the Objective Multidimensional

Assessment of Youth

LEBER, W. R. (University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA)

**Clinical Interviewing*

LICHTENBERGER, E. O. (The Salk Institute, La Jolla, CA, USA)

**Intellectual Assessment*

MARUISH, M. E. (Strategic Advantage Inc., Minneapolis, MN, USA)

Therapeutic Assessment: Linking Assessment and Treatment

OLLENDICK, T. H. (Virginia Tech, Blacksburg, VA, USA)

**Principles and Practices of Behavioral Assessment with Children*

REYNOLDS, C. R. (Texas A&M University, College Station, TX, USA)

*Fundamentals of Measurement and Assessment in Psychology; *Neuropsychological Assessment of Children*

RICCIO, C. A. (Texas A&M University, College Station, TX, USA)

**Neuropsychological Assessment of Children*

ROID, G. H. (George Fox University, Newberg, OR , USA)

**Computer Assisted Psychological Assessment*

SECHREST, L. (University of Arizona, Tucson, AZ, USA)

**The Role of Assessment in Clinical Psychology*

STEWART, M. (University of Arizona, Tucson, AZ, USA)

**The Role of Assessment in Clinical Psychology*

STICKLE, T. R. (University of Arizona, Tucson, AZ, USA)

**The Role of Assessment in Clinical Psychology*

TAYLOR, J. (University of Minnesota, Minneapolis, MN, USA)

**Objective Personality Assessment with Adults*

TEGLASI, H. (University of Maryland, College Park, MD, USA)

Assessment of Schema and Problem-solving Strategies with Projective Techniques

TROBST, K. K. (University of British Columbia, Vancouver, BC, Canada)

**Principles of Personality Assessment*

WEINER, I. B. (University of South Florida, Tampa, FL, USA)

**Projective Assessment of Children and Adolescents*

WIGGINS, J. S. (University of British Columbia, Vancouver, BC, Canada)

**Principles of Personality Assessment*

WILLIS, D. J. (University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA)

**Clinical Interviewing*

4.01

The Role of Assessment in Clinical Psychology

LEE SECHREST, TIMOTHY R. STICKLE, and MICHELLE STEWART
University of Arizona, Tucson, AZ, USA

4.01.1 INTRODUCTION	2
4.01.1.1 <i>Useful Clinical Assessment is Difficult but not Impossible</i>	2
4.01.2 WHY ARE ASSESSMENTS DONE?	4
4.01.2.1 <i>Bounded vs. Unbounded Inference and Prediction</i>	5
4.01.2.2 <i>Prevalence and Incidence of Assessment</i>	5
4.01.2.3 <i>Proliferation of Assessment Devices</i>	7
4.01.2.4 <i>Over-reliance on Self-report</i>	9
4.01.3 PSYCHOMETRIC ISSUES WITH RESPECT TO CURRENT MEASURES	10
4.01.3.1 <i>Reliability</i>	10
4.01.3.2 <i>Validity</i>	10
4.01.3.3 <i>Item Response Theory</i>	11
4.01.3.4 <i>Scores on Tests</i>	11
4.01.3.5 <i>Calibration of Measures</i>	12
4.01.4 WHY HAVE WE MADE SO LITTLE PROGRESS?	12
4.01.4.1 <i>The Absence of the Autopsy</i>	14
4.01.5 FATEFUL EVENTS CONTRIBUTING TO THE HISTORY OF CLINICAL ASSESSMENT	14
4.01.5.1 <i>The Invention of the Significance Test</i>	14
4.01.5.2 <i>Ignoring Decision Making</i>	14
4.01.5.3 <i>Seizing on Construct Validity</i>	16
4.01.5.4 <i>Adoption of the Projective Hypothesis</i>	16
4.01.5.5 <i>The Invention of the Objective Test</i>	17
4.01.5.6 <i>Disinterest in Basic Psychological Processes</i>	17
4.01.6 MISSED SIGNALS	19
4.01.6.1 <i>The Scientist–Practitioner Model</i>	19
4.01.6.2 <i>Construct Validity</i>	19
4.01.6.3 <i>Assumptions Underlying Assessment Procedures</i>	20
4.01.6.4 <i>Antecedent Probabilities</i>	20
4.01.6.5 <i>Need for Integration of Information</i>	20
4.01.6.6 <i>Method Variance</i>	21
4.01.6.7 <i>Multiple Measures</i>	21
4.01.7 THE ORIGINS OF CLINICAL ASSESSMENT	22
4.01.7.1 <i>The Tradition of Assessment in Psychology</i>	22
4.01.7.1.1 <i>Witmer</i>	23
4.01.7.1.2 <i>Army Alpha</i>	23
4.01.8 THE RORSCHACH INKBLOT TECHNIQUE AND CLINICAL PSYCHOLOGY	23
4.01.8.1 <i>The Social and Philosophical Context for the Appearance of the Rorschach</i>	23
4.01.8.2 <i>The Birth of the Rorschach</i>	24
4.01.8.3 <i>Clinical vs. Statistical Prediction</i>	25
4.01.8.4 <i>Old Tests Never Die, They Just Fade Away</i>	26

4.01.9 OTHER MEASURES USED IN CLINICAL PSYCHOLOGY	27
4.01.9.1 <i>The Thematic Apperception Test</i>	27
4.01.9.2 <i>Sentence Completion Tests</i>	28
4.01.9.3 <i>Objective Testing</i>	28
4.01.9.4 <i>The Clinician as a Clinical Instrument</i>	28
4.01.9.5 <i>Structured Interviews</i>	29
4.01.10 CONCLUSIONS	29
4.01.11 REFERENCES	29

4.01.1 INTRODUCTION

In this chapter we will describe the current state of affairs with respect to assessment in clinical psychology and then we will attempt to show how clinical psychology got to that state, both in terms of positive influences on the directions that efforts in assessment have taken and in terms of missed opportunities for alternative developments that might have been more productive psychology. For one thing, we really do not think the history is particularly interesting in its own right. The account and views that we will give here are our own; we are not taking a neutral—and innocuous—position. Readers will not find a great deal of equivocation, not much in the way of “a glass half-empty is, after all, half-full” type of placation. By assessment in this chapter, we refer to formal assessment procedures, activities that can be named, described, delimited, and so on. We assume that all clinical psychologists are more or less continuously engaged in informal assessment of clients with whom they work. Informal assessment, however, does not follow any particular pattern, involves no rules for its conduct, and is not set off in any way from other clinical activities. We have in mind assessment procedures that would be readily defined as such, that can be studied systematically, and whose value can be quantified. We will not be taking account of neuropsychological assessment nor of behavioral assessment, both of which are covered in other chapters in this volume. It will help, we think, if we begin by noting the limits within which our critique of clinical assessment is meant to apply. We, ourselves, are regularly engaged in assessment activities, including development of new measures, and we are clinicians, too.

4.01.1.1 Useful Clinical Assessment is Difficult but not Impossible

Many of the comments about clinical assessment that follow may seem to some readers to be pessimistic and at odds with the experiences of professional clinicians. We think our views are quite in accord with both research and the

theoretical underpinnings for assessment activities, but in at least some respects we are not so negative in our outlook as we may seem. Let us explain. In general, tests and related instruments are devised to measure constructs, for example, intelligence, ego strength, anxiety, antisocial tendencies. In that context, it is reasonable to focus on the construct validity of the test at hand: how well does the test measure the construct it is intended to measure? Generally speaking, evaluations of tests for construct validity do not produce single quantitated indexes. Rather, evidence for construct validity consists of a “web of evidence” that fits together at least reasonably well and that persuades a test user that the test does, in fact, measure the construct at least passably well. The clinician examiner especially if he or she is acquainted in other ways with the examinee, may form impressions, perhaps compelling, of the validity of test results. The situation may be something like the following:

Test <—construct

That is, the clinician uses a test that is a measure of a construct. The path coefficient relating the test to the construct (in the convention of structural equations modeling, the construct causes the test performance) may well be substantial. A more concrete example is provided by the following diagram:

IQ Test <—0.80—intelligence

This diagram indicates that the construct of intelligence causes performance on an IQ test. We believe that IQ tests may actually be quite good measures of the construct of “intelligence.” Probably clinicians who give intelligence tests believe that in most instances the test gives them a pretty good estimate of what we mean by intelligence, for example, 0.80 in this example. To use a term that will be invoked later, the clinician is “enlightened” by the results from the test.

As long as the clinical use of tests is confined to enlightenment about constructs, many tests may have reasonably good, maybe even very good “validity.” The tests are good measures of the constructs. In many, if not most, clinical uses of tests, however, the tests are used in order to make decisions. Tests are used, for example to

decide whether a parent should have custody of a child, to decide whether a patient is likely to benefit from some form of therapy, to decide whether a child “should” be placed in a social classroom, or to decide whether a patient should be put on some particular medication. Using our IQ test example, we get a diagram of the following sort:

IQ Test <—0.80—intelligence—0.50—>
School grades

This diagram, which represents prediction rather than simply enlightenment, has two paths, and the second path is almost certain to have a far lower validity coefficient than the first one. Intelligence has a stronger relationship to performance on an IQ test than to performance in school. If an IQ test had construct validity of 0.80, and if intelligence as a construct were correlated 0.50 with school grades, which means that intelligence would account for 25% of the total variance in school grades, then the correlation between the IQ test and school grades would be only $0.80 \times 0.50 = 0.40$ (which is about what is generally found to be the case).

IQ Test <—0.40—> School grades

A very good measure of ego strength may not be a terribly good predictor of resistance to stress in some particular set of circumstances. Epstein (1983) pointed out some time ago that tests cannot be expected to be related especially well to specific behaviors, but it is in relation to specific behaviors that tests are likely to be used in clinical settings.

It could be argued and has been, (e.g., Meyer & Handler 1997), that even modest validities like 0.40 are important. Measures with a validity of 0.40, for example, can improve one's prediction from that 50% of a group of persons will succeed at some task to the prediction that 70% will succeed. If the provider of a service cannot serve all eligible or needy persons, that improvement in prediction may be quite useful.

In clinical settings, however, decisions are made about individuals, not groups. To recommend that one person should not receive a service because the chances of benefit from the service are only 30% instead of the 50% that would be predicted without a test, could be regarded as a rather bold decision for a clinician to make about a person in need of help. Hunter and Schmidt (1990) have developed very useful approaches to validity generalization that usually result in estimates of test validity well above the correlations reported in actual use, but their estimates apply at the level of theory, construct validity, rather than at the level of specific application as in clinical settings.

A recommendation to improve the clinical uses of tests can actually be made: test for more things. Think of the determinants of perfor-

mance in school, say college, as an example. College grades depend on motivation, persistence, physical health, mental health, study habits, and so on. If clinical psychologists are serious about predicting performance in college, then they probably will need to measure several quite different constructs and then combine all those measures into a prediction equation. The measurement task may seem onerous, but it is worth remembering Cronbach's (1960) bandwidth vs. fidelity argument: it is often better to measure more things less well than to measure one thing extraordinarily well. A lot of measurement could be squeezed into the times usually allotted to low bandwidth tests. The genius of the profession will come in the determination of what to measure and how to measure it. The combination of all the information, however, is likely best to be done by a statistical algorithm for reasons that we will show later.

We are not negative toward psychological testing, but we think it is a lot more difficult and complicated than it is generally taken to be in practice. An illustrative case is provided by the differential diagnosis of attention deficit hyperactivity disorder (ADHD). There might be an ADHD scale somewhere but a more responsible clinical study would recognize that the diagnosis can be difficult, and that the validity and certainty of the diagnosis of ADHD is greatly improved by using multiple measures and multiple reporting agents across multiple contexts. For example, one authority recommended beginning with an initial screening interview, in which the possibility of an ADHD diagnosis is ruled in, followed by an extensive assessment battery addressing multiple domains and usually including (depending upon age): a Wechsler Intelligence Scale for Children (WISC-III; McCracken & McCallum, 1993), a behavior checklist (e.g., Youth Self-Report (YSR); Achenbach & Edelbrock, 1987), an academic achievement battery (e.g., Kaufmann Assessment Battery for Children; Kaufmann & Kaufmann, 1985), a personality inventory (e.g., Millon Adolescent Personality Inventory (MAPI); Millon & Davis, 1993), a computerized sustained attention and distractibility test (Gordon Diagnostic System [GDS]; McClure & Gordon, 1984), and a semistructured or a structured clinical interview (e.g., Diagnostic Interview Schedule for Children [DISC]; Costello, Edelbrock, Kalas, Kessler, & Klaric, 1982).

The results from the diagnostic assessment may be used to further rule in or rule out ADHD as a diagnosis, in conjunction with child behavior checklists (e.g., CBCL, Achenbach & Edelbrock, 1983; Teacher Rating Scales, Goyette, Conners, & Ulrich, 1978), completed by the parent(s) and teacher, and additional

school performance information. The parent and teacher complete both a historical list and then a daily behavior checklist for a period of two weeks in order to adequately sample behaviors. The information from home and school domains may be collected concurrently with evaluation of the diagnostic assessment battery, or the battery may be used initially to continue to rule in the diagnosis as a possibility, and then proceed with collateral data collection. We are impressed with the recommended ADHD diagnostic process, but we do recognize that it would involve a very extensive clinical process that would probably not be reimbursable under most health insurance plans. We would also note, however, that the overall diagnostic approach is not based on any decision-theoretic approach that might guide the choice of instruments corresponding to a process of decision making. Or alternatively, the process is not guided by any algorithm for combining information so as to produce a decision. Our belief is that assessment in clinical psychology needs the same sort of attention and systematic study as is occurring in medical areas through such organizations as the Society for Medical Decision Making.

In summary, we think the above scenario, or similar procedures using similar instruments (e.g., Atkins, Pelham, & White, 1990; Hoza, Vollano, & Pelham, 1995), represent an exemplar of assessment practice. It should be noted, however, that the development of such multimodal batteries is an iterative process. One will soon reach the point of diminishing returns in the development of such batteries, and the incremental validity (Sechrest, 1963) of instruments should be assessed. ADHD is an example in which the important domains of functioning are understood, and thus can be assessed. We know of no examples other than ADHD of such systematic approaches to assessment for decision making. Although approaches such as described here and by Pelham and his colleagues appear to be far from standard practice in the diagnosis of ADHD, we think they ought to be. The outlined procedure is modeled after a procedure developed by Gerald Peterson, Ph.D., Institute for Motivational Development, Bellevue, WA.

4.01.2 WHY ARE ASSESSMENTS DONE?

Why do we “test” in the first place? It is worth thinking about all the instances in which we do not test. For example, we usually do not test our own children—nor our spouses. That is because we have ample opportunities to observe the “performances” in which we are interested. That

may be one reason that psychotherapists are disinclined to test their own clients: they have many opportunities to observe the behaviors in which they are interested, that is, if not the actual behaviors than reasonably good indicators of them. As we see it, testing is done primarily for one or more of three reasons: efficiency of observation, revealing cryptic conditions, and quantitative tagging.

Testing may provide for more efficient observation than most alternatives. For example, “tailing” a person, that method so dear to detective story writers, would prove definitive for many dispositions, but it would be expensive and often impractical or even unethical (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Testing may provide for more efficient observation than most alternatives. It seems unlikely that any teacher would not have quite a good idea of the intelligence and personality of any of her pupils after at most a few weeks of a school year, but appropriate tests might provide useful information from the very first day. Probably clinicians involved in treating patients do not anticipate much gain in useful information after having held a few sessions with a patient. In fact, they may not anticipate much gain under most circumstances, which could account for the apparent infrequent use of assessment procedures in connection with psychological treatment.

Testing is also done in order to uncover “cryptic” conditions, that is, characteristics that are hidden from view or otherwise difficult to discern. In medicine, for example, a great many conditions are cryptic, blood pressure being one example. It can be made visible only by some device. Cryptic conditions have always been of great interest in clinical psychology, although their importance may have been exaggerated considerably. The Rorschach, a prime example of a putative decrypter, was hailed upon its introduction as “providing a window on the mind,” and it was widely assumed that in skillful hands the Rorschach would make visible a wide range of hidden dispositions, even those unknown to the respondent (i.e., in “the unconscious”). Similarly, the Thematic Apperception Test was said to “expose underlying inhibited tendencies” of which the subject is unaware and to permit the subject to leave the test “happily unaware that he has presented the psychologist with what amounts to an X-ray picture of his inner self” (Murray, 1943, p. 1).

Finally, testing may be done, is often done, in order to provide a quantitative “tag” for some dispositions or other characteristic. In foot races, to take a mundane example, no necessity exists to time the races; it is sufficient to determine simply the order of the finish.

Nonetheless, races are timed so that each one may be quantitatively tagged for sorting and other uses, for example, making comparisons between races. Similarly, there is scarcely ever any need for more than a crude indicator of a child's intelligence, for example, "well above average," such as a teacher might provide. Nonetheless, the urge to seemingly precise quantification is strong, even if the precision is specious, and tests are used regularly to provide such estimates as "at the 78th percentile in aggression" or "IQ = 118." Although quantitative tags are used, and may be necessary, for some decision-making, for example, the awarding of scholarships based on SAT scores, it is to be doubted that such tags are ever of much use in clinical settings.

4.01.2.1 Bounded vs. Unbounded Inference and Prediction

Bounded prediction is the use of a test or measure to make some limited inference or prediction about an individual, couple, or family, a prediction that might be limited in time, situation, or range of behavior (Levy, 1963; Sechrest, 1968). Some familiar examples of bounded prediction are that of a college student's grade point average based on their SAT score, assessing the likely response of an individual to psychotherapy for depression based on MMPI scores and a SCID interview, or prognosticating outcome for a couple in marital therapy given their history. These predictions are bounded because they are using particular measures to predict a specified outcome in a given context. Limits to bounded predictions are primarily based on knowledge of two areas. First, the reliability of the information, that is, interview or test, for the population from which the individual is drawn. Second, and most important, these predictions are based on the relationship between the predictor and the outcome. That is to say, they are limited by the validity of the predictor for the particular context in question.

Unbounded inference or prediction, which is common in clinical practice, is the practice of making general assessment of an individual's tendencies, dispositions, and behavior, and inferring prognosis for situations that may not have been specified at the time of assessment. These are general statements made about individuals, couples, and families based on interviews, diagnostic tests, response to projective stimuli, and so forth that indicate how these people are likely to behave across situations. Some unbounded predictions are simply descriptive statements, for example, with respect to personality, from which at some future time the

clinician or another person might make an inference about a behavior not even imagined at the time of the original assessment. A clinician might be asked to apply previously obtained assessment information to an individual's ability to work, ability as a parent, likelihood of behaving violently, or even the probability that an individual might have behaved in some way in the past (e.g., abused a spouse or child). Thus, they are unbounded in context. Since reliability and validity require context, that is, a measure is reliable in particular circumstances, one cannot readily estimate the reliability and validity of a measure for unspecified circumstances.

To the extent that the same measures are used repeatedly to make the same type of prediction or judgment about individuals, the more the prediction becomes of a bounded nature. Thus, an initially unbounded prediction becomes bounded by the consistency of circumstances of repeated use. Under these circumstances, reliability, utility, and validity can be assessed in a standard manner (Sechrest, 1968). Without empirical data, unbounded predictions rest solely upon the judgment of the clinician, which has proven problematic (see Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996; Meehl, 1954). Again, the contrast with medical testing is instructive. In medicine, tests are generally associated with gathering additional information about specific problems or systems. Although one might have a "wellness" visit to detect level of functioning and signs of potential problems, it would be scandalous to have a battery of medical tests to "see how your health might be" under an unspecified set of circumstances. Medical tests are bounded. They are for specific purposes at specific times.

4.01.2.2 Prevalence and Incidence of Assessment

It is interesting to speculate about how much assessment is actually done in clinical psychology today. It is equally interesting to realize how little is known about how much assessment is done in clinical psychology today. What little is known has to do with "incidence" of assessment, and that only from the standpoint of the clinician and only in summary form. Clinical psychologists report that a modest amount of their time is taken up by assessment activities. The American Psychological Association's (APA's) Committee for the Advancement of Professional Practice (1996) conducted a survey in 1995 of licensed APA members. With a response rate of 33.8%, the survey suggested that psychologists spend about 14% of their time conducting assessments—roughly six or seven hours per week. The low response rate, which ought to be considered disgraceful in a

profession that claims to survive by science, is indicative of the difficulties involved in getting useful information about the practice of psychology in almost any area. The response rate was described as "excellent" in the report of the survey. Other estimates converge on about the same proportion of time devoted to assessment (Wade & Baker, 1977; Watkins, 1991; Watkins, Campbell, Nieberding, & Hallmark, 1995). Using data across a sizable number of surveys over a considerable period of time, Watkins (1991) concludes that about 50–75% of clinical psychologists provide at least some assessment services. We will say more later about the relative frequency of use of specific assessment procedures, but Watkins et al. (1995) did not find much difference in relative use across seven diverse work settings.

Think about what appears not to be known: the number of psychologists who do assessments in any period of time; the number of assessments that psychologists who do them actually do; the number or proportion of assessments that use particular assessment devices; the proportion of patients who are subjected to assessments; the problems for which assessments are done. And that does not exhaust the possible questions that might be asked. If, however, we take seriously the estimate that psychologists spend six or seven hours per week on assessment, then it is unlikely that those psychologists who do assessments could manage more than one or two per week; hence, only a very small minority of patients being seen by psychologists could be undergoing assessment. Wade and Baker (1977) found that psychologists claimed to be doing an average of about six objective tests and three projective tests per week, and that about a third of their clients were given at least one or the other of the tests, some maybe both. Those estimates do not make much sense in light of the overall estimate of only 15% of time (6–8 hours) spent in testing.

It is almost certain that those assessment activities in which psychologists do engage are carried out on persons who are referred by some other professional person or agency specifically for assessment. What evidence exists indicates that very little assessment is carried out by clinical psychologists on their own clients, either for diagnosis or for planning of treatment. Nor is there any likelihood that clinical psychologists refer their own clients to some other clinician for assessment. Some years ago, one of us (L. S.) began a study, never completed, of referrals made by clinical psychologists to other mental health professionals. The study was never completed in part because referrals were, apparently, very infrequent, mostly having to do with troublesome patients. A total of about

40 clinicians were queried, and in no instance did any of those clinical psychologists refer any client for psychological assessment.

Thus, we conclude that only a small minority of clients or patients of psychologists are subjected to any formal assessment procedures, a conclusion supported by Wade and Baker (1977) who found that relatively few clinicians appear to use standard methods of administration and scoring. Despite Wade and Baker's findings, it also seems likely that clinical psychologists do very little assessment on their own clients. Most assessments are almost certainly on referral. Now contrast that state of affairs with the practice of medicine: assessment is at the heart of medical practice. Scarcely a medical patient ever gets any substantial treatment without at least some assessment. Merely walking into a medical clinic virtually guarantees that body temperature and blood pressure will be measured. Any indication of a problem that is not completely obvious will result in further medical tests, including referral of patients from the primary care physician to other specialists.

The available evidence also suggests that psychologists do very little in the way of formal assessment of clients prior to therapy or other forms of intervention. For example, books on psychological assessment even in clinical psychology may not even mention psychotherapy or other interventions (e.g., see Maloney & Ward, 1976), and the venerated and authoritative *Handbook of psychotherapy and behavior change* (Bergen & Garfield, 1994) does not deal with assessment except in relation to diagnosis and the prediction of response to therapy and to determining the outcomes of therapy, that is, there is no mention of assessment for planning therapy at any stage in the process. That is, we think, anomalous, especially when one contemplates the assessment activities of other professions. It is almost impossible even to get to speak to a physician without at least having one's temperature and blood pressure measured, and once in the hands of a physician, almost all patients are likely to undergo further explicit assessment procedures, for example, auscultation of the lungs, heart, and carotid arteries. Unless the problem is completely obvious, patients are likely to undergo blood or other body-fluid tests, imaging procedures, assessments of functioning, and so on. The same contrast could be made for chiropractors, speech and hearing specialists, optometrists, and, probably, nearly all other clinical specialists. Clinical psychology appears to have no standard procedures, not much interest in them, and no instruments for carrying them out in any case. Why is that?

One reason, we suspect, is that clinical psychology has never shown much interest in normal functioning and, consequently, does not have very good capacity to identify normal responses or functioning. A competent specialist in internal medicine can usefully palpate a patient's liver, an organ he or she cannot see, because that specialist has been taught what a normal liver should feel like and what its dimensions should (approximately) be. A physician knows what normal respiratory sounds are. An optometrist certainly knows what constitutes normal vision and a normal eye. Presumably, a chiropractor knows a normal spine when he or she sees one.

Clinical psychology has no measures equivalent to body temperature and blood pressure, that is, quick, inexpensive screeners (vital signs) that can yield "normal" as a conclusion just as well as "abnormal." Moreover, clinical psychologists appear to have a substantial bias toward detection of psychopathology. The consequence is that clinical psychological assessment is not likely to provide a basis for a conclusion that a given person is "normal," and that no intervention is required. Obviously, the case is different for "intelligence," for which the conclusion of "average" or some such is quite common.

By their nature, psychological tests are not likely to offer many surprises. A medical test may reveal a completely unexpected condition of considerable clinical importance, for example, even in a person merely being subjected to a routine examination. Most persons who come to the attention of psychologists and other mental health professionals are there because their behavior has already betrayed important anomalies, either to themselves or to others. A clinical psychologist would be quite unlikely to administer an intelligence test to a successful business man and discover, completely unexpectedly, that the man was really "stupid." Tests are likely to be used only for further exploration or verification of problems already evident. If they are already evident, then the clinician managing the case may not see any particular need for further assessment.

A related reason that clinical psychologists appear to show so little inclination to do assessment of their own patients probably has to do with the countering inclination of clinical psychologists, and other similarly placed clinicians, to arrive at early judgments of patients based on initial impressions. Meehl (1960) noted that phenomenon many years ago, and it likely has not changed. Under those circumstances, testing of clients would have very little incremental value (Sechrest, 1963) and would seem unnecessary. At this point, it may be worth repeating that apparently no information is

available on the specific questions for which psychologists make assessments when they do so.

Finally, we do believe that current limitations on practice imposed by managed care organizations are likely to limit even further the use of assessment procedures by psychologists. Pressures are toward very brief interventions, and that probably means even briefer assessments.

4.01.2.3 Proliferation of Assessment Devices

Clinical psychology has experienced an enormous proliferation of tests since the 1960s. We are referring here to commercially published tests, available for sale and for use in relation to clinical problems. For example, inspection of four current test catalogs indicates that there are at least a dozen different tests (scales, inventories, checklists, etc.) related to attention deficit disorder (ADD) alone, including forms of ADD that may not even exist, for example, adult ADD. One of the test catalogs is 100 pages, two are 176 pages, and the fourth is an enormous 276 pages. Even allowing for the fact that some catalog pages are taken up with advertisements for books and other such, the amount of test material available is astonishing. These are only four of perhaps a dozen or so catalogs we have in our files.

In the mid-1930s Buros published the first listings of psychological tests to help guide users in a variety of fields in choosing an appropriate assessment instrument. These early uncritical listings of tests developed into the *Mental measurements yearbook* and by 1937 the listings had expanded to include published test reviews. The *Yearbook*, which includes tests and reviews of new and revised tests published for commercial use, has continued to grow and is now in its 12th edition (1995). The most recent edition reviewed 418 tests available for use in education, psychology, business, and psychiatry. *Buros Mental Measurements Yearbook* is a valuable resource for testers, but it also charts the growth of assessment instruments. In addition to instruments published for commercial use, there are scores of other tests developed yearly for noncommercial use that are never reviewed by Buros. Currently, there are thousands of assessment instruments available for researchers and practitioners to choose from.

The burgeoning growth in the number of tests has been accompanied by increasing commercialization as well. The monthly *Monitor* published by the APA is replete with ads for test instruments for a wide spectrum of purposes. Likewise, APA conference attendees are inundated with preconference mailings advertising tests and detailing the location of

the test publisher's booth at the conference site. Once at the conference, attendees are often struck by the slick presentation of the booths and hawking of the tests. Catalogs put out by test publishers are now also slick, in more ways than one. They are printed in color on coated paper and include a lot of messages about how convenient and useful the tests are with almost no information at all about reliability and validity beyond assurances that one can count on them.

The proliferation of assessment instruments and commercial development are not inherently detrimental to the field of clinical psychology. They simply make it more difficult to choose an appropriate test that is psychometrically sound, as glib ads can be used as a substitute for the presentation of sound psychometric properties and critical reviews. This is further complicated by the availability of computer scoring and software that can generate assessment reports. The ease of computer-based applications such as these can lead to their uncritical application by clinicians. Intense marketing of tests may contribute to their misuse, for example, by persuading clinical psychologists that the tests are remarkably simple and by convincing those same psychologists that they know more than they actually do about tests and their appropriate uses.

Multiple tests, even several tests for every construct, might not necessarily be a bad idea in and of itself, but we believe that the resources in psychology are simply not sufficient to support the proper development of so many tests. Few of the many tests available can possibly be used on more than a very few thousand cases per year, and perhaps not even that. The consequence is that profit margins are not sufficient to support really adequate test development programs. Tests are put on the market and remain there with small normative samples, with limited evidence for validity, which is much more expensive to produce than evidence for reliability, and with almost no prospect for systematic exploration of the other psychometric properties of the items, for example, discrimination functions or tests of their calibration (Sechrest, McKnight, & McKnight, 1996).

One of us (L. S.) happens to have been a close spectator of the development of the SF-36, a now firmly established and highly valued measure of health and functional status (Ware & Sherbourne, 1992). The SF-36 took 15–20 years for its development, having begun as an item pool of more than 300 items. Over the years literally millions of dollars were invested in the development of the test, and it was subjected, often repeatedly, to the most sophisticated psychometric analyses and to detailed scrutiny

of every individual item. The SF-36 has now been translated into at least 37 languages and is being used in an extraordinarily wide variety of research projects. More important, however, the SF-36 is also being employed routinely in evaluating outcomes of clinical medical care. Plans are well advanced for use of the SF-36 that will result in its administration to 300 000 patients in managed care every year. It is possible that over the years the Wechsler intelligence tests might have a comparable history of development, and the Minnesota Multiphasic Inventory (MMPI) has been the focus of a great many investigations, as has the Rorschach. Neither of the latter, however, has been the object of systematic development efforts funded centrally, and scarcely any of the many other tests now available are likely to be subjected to anything like the same level of development effort (e.g., consider that in its more than 70-year history, the Rorschach has never been subjected to any sort of revision of its original items).

Several factors undoubtedly contribute to the proliferation of psychological tests (not the least, we suspect, being their eponymous designation and the resultant claim to fame), but surely one of the most important would be the fragmentation of psychological theory, or what passes for theory. In 1995 a taskforce was assembled under the auspices of the APA to try to devise a uniform test (core) battery that would be used in all psychotherapy research studies (Strupp, Horowitz, & Lambert, 1997). The effort failed, in large part because of the many points of view that seemingly had to be represented and the inability of the conferees to agree even on any outcomes that should be common to all therapies. Again, the contrast with medicine and the nearly uniform acceptance of the SF-36 is stark.

Another reason for the proliferation of tests in psychology is, unquestionably, the seeming ease with which they may be "constructed." Almost anyone with a reasonable "construct" can write eight or 10 self-report items to "measure" it, and most likely the new little scale will have "acceptable" reliability. A correlation or two with some other measure will establish its "construct validity," and the rest will eventually be history. All that is required to establish a new projective test, it seems, is to find a set of stimuli that have not, according to the published literature, been used before and then show that responses to the stimuli are suitably strange, perhaps stranger for some folks than others. For example, Sharkey and Ritzler (1985) noted a new Picture Projective Test that was created by using photographs from a photo essay. The pictures

were apparently selected based on the authors' opinions about their ability to elicit "meaningful projective material," meaning responses with affective content and activity themes. No information was given pertaining to comparison of various pictures and their responses nor relationships to other measures of the target constructs; no comparisons were made to pictures that were deemed inappropriate. The "validation" procedure simply compared diagnoses to those in charts and results of the TAT. Although rater agreement was assessed, there was no formal measurement of reliability.

New tests are cheap, it seems. One concern is that so many new tests appear also to imply new constructs, and one wonders whether clinical psychology can support anywhere near as many constructs as are implied by the existence of so many measures of them. Craik (1986) made the eminently sensible suggestion that every "new" or infrequently used measure used in a research project should be accompanied by at least one well-known and widely used measure from the same or a closely related domain. New measures should be admitted only if it is clear that they measure something of interest and are not redundant, that is, have discriminant validity. That recommendation would likely have the effect of reducing the array of measures in clinical psychology by remarkable degrees if it were followed.

The number of tests that are taught in graduate school for clinical psychology is far lower than the number available for use. The standard stock-in-trade are IQ tests such as the Wechsler Adult Intelligence Scale (WAIS), personality profiles such as the MMPI, diagnostic instruments (Structured Clinical Interview for *DSM-III-R* [SCID]), and at some schools, the Rorschach as a projective test. This list is rounded out by a smattering of other tests like the Beck Depression Inventory and Millon. Recent standard application forms for clinical internships developed by the Association of Psychology Postdoctoral and Internship Centers (APPIC) asked applicants to report on their experience with 47 different tests and procedures used for adult assessment and 78 additional tests used with children! It is very doubtful that training programs actually provide training in more than a handful of the possible devices.

Training in testing (assessment) is not at all the same as training in measurement and psychometrics. Understanding how to administer a test is useful but cannot substitute for evaluating the psychometric soundness of tests. Without grounding in such principles, it is easy to fall prey to glib ads and ease of computer administration without questioning the quality

of the test. Psychology programs appear, unfortunately, to be abandoning training in basic measurement and its theory (Aiken, West, Sechrest, & Reno, 1990).

4.01.2.4 Over-reliance on Self-report

"Where does it hurt?" is a question often heard in physicians' offices. The physician is asking the patient to self-report on the subjective experience of pain. Depending on the answer, the physician may prescribe some remedy, or may order tests to examine the pain more thoroughly and obtain objective evidence about the nature of the affliction before pursuing a course of treatment. The analog heard in psychologists' offices is "How do you feel?" Again, the inquiry calls forth self-report on a subjective experience and like the physician, the psychologist may determine that tests are in order to better understand what is happening with the client.

When the medical patient goes for testing, she or he is likely to be poked, prodded, or pricked so that blood samples and X-rays can be taken. The therapy client, in contrast, will most likely be responding to a series of questions in an interview or answering a pencil-and-paper questionnaire. The basic difference between these is that the client in clinical psychology will continue to use self-report in providing a sample, whereas the medical patient will provide objective evidence.

Despite the proliferation of tests in recent years, few rely on evidence other than the client's self-report for assessing behavior, symptoms, or mood state. Often assessment reports remark that the information gleaned from testing was corroborated by interview data, or vice versa, without recognizing that both rely on self-report alone. The problems with self-report are well documented: poor recall of past events, motivational differences in responding, social desirability bias, and malinger, for example. Over-reliance on self-report is a major criticism of psychological assessment as it is currently conducted and was the topic of a recent conference sponsored by the National Institute of Mental Health.

What alternatives are there to self-report? Methods of obtaining data on a client's behavior that do not rely on self-report do exist. Behavioral observation with rating by judges can permit the assessment of behavior, often without the client's awareness or outside the confines of an office setting. Use of other informants such as family members or co-workers to provide data can yield valuable information about a client. Yet, all too often these alternatives are not pursued because they involve time or resources—in short, they are

demanding approaches. Compared with asking a client about his or her mood state over the last week, organizing field work or contacting informants involves a great deal more work and time.

Instruments are available to facilitate collection of data not relying so strongly on self-report and for collection of data outside the office setting, for example, the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1983). The CBCL is meant to assist in diagnosing a range of psychological and behavior problems in children, and it relies on parent, teacher, and self-reports of behavior. Likewise, neuropsychological tests utilize functional performance measures much more than self-report. However, as Craik (1986) noted with respect to personality research, methods such as field studies are not widely used as alternatives to self-report. This problem of over-reliance on self-report is not new (see Webb, Campbell, Schwartz, & Sechrest, 1966).

4.01.3 PSYCHOMETRIC ISSUES WITH RESPECT TO CURRENT MEASURES

Consideration of the history and current status of clinical assessment must deal with some fundamental psychometric issues and practices. Although psychometric is usually taken to refer to reliability and validity of measures, matters are much more complicated than that, particularly in light of developments in psychometric theory and method since the 1960s, which seem scarcely to have penetrated clinical assessment as an area. Specifically, generalizability theory and Item Response Theory (IRT) offer powerful tools with which to explore and develop clinical assessment procedures, but they have seen scant use in that respect.

4.01.3.1 Reliability

The need for “reliable” measures is by now well accepted in all of psychology, including clinical assessment. What is not so widespread is the necessary understanding of what constitutes reliability and the various uses of that term. In their now classic presentation of generalizability theory, Cronbach and his associates (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) used the term “dependability” in a way that is close to what is meant by reliability, but they made especially clear, as classical test theory had not, that measures are dependable (generalizable) in very specific ways, that is, that they are dependable across some particular conditions of use (facets), and assessments of dependability are not at all interchangeable. For example, a

given assessment may be highly dependable across particular items but not necessarily across time. An example might be a measure of mood, which ought to have high internal consistency (i.e., across items) but that might not, in fact, should not, have high dependability over time, else the measure would be better seen as a trait rather than as a mood measure.

An assessment procedure might be highly dependable in terms of internal consistency and across time but not satisfactorily dependable across users, for example, being susceptible to a variety of biases characteristic of individual clinicians. Or an assessment procedure might not be adequately dependable across conditions of its use, as might be the case when a measure is taken from a research to a clinical setting. Or an assessment procedure might not be dependable across populations, for example, a projective instrument useful with mental patients might be misleading if used with imaginative and playful college students.

Issues of dependability are starkly critical when one notes the regrettably common practice of justifying the use of a measure on the ground that it is “reliable,” often without even minimal specification of the facet(s) across which that reliability was established. The practice is even more regrettable when, as is often the case, only a single value for reliability is given when many are available and when one suspects that the figure reported was not chosen randomly from those available. Moreover, it is all too frequently the case that the reliability estimate reported is not directly relevant to the decisions to be made. Internal consistency, for example, may not be as important as generalizability over time when one is using a screening instrument. That is, if one is screening in a population for psychopathology, it may not be of great interest that two persons with the same scores are different in terms of their manifestations of pathology, but it is of great interest whether if one retested them a day or so later, the scores would be roughly consistent.

In short, clinical assessment in psychology is unfortunately casual in its use of reliability estimates, and it is shamefully behind the curve in its attention to the advantages provided by generalizability theory, originally proposed in 1963 (Cronbach, Rajaratnam, & Gleser, 1963).

4.01.3.2 Validity

It is customary to treat validity of measures as a topic separate from reliability, but we think that is not only unnecessary but undesirable. In our view, the validity of measures is simply an extension of generalizability theory to the question of what other performances aside from

those involved in the test is the score generalizable. A test score that is generalizable to another very similar performance, say on the same set of test items or over a short period of time, is said to be reliable. A test score that is generalizable to a score on another similar test is sometimes said to be “valid,” but we think that a little reflection will show that unless the tests demand very different kinds of performances, generalizability from one test to another is not much beyond the issues usually regarded as having to do with reliability. When, however, a test produces a score that is informative about another very different kind of performance, we gradually move over into the realm termed validity, such as when a paper-and-pencil test of “readiness for change” (Prochaska, DiClemente, & Norcross, 1992) predicts whether a client will benefit from treatment or even just stay in treatment.

We will say more later about construct validity, but a test or other assessment procedure may be said to have construct validity if it produces generalizable information and if that information relates to performances that are conceptually similar to those implied by the name or label given to the test. Essentially, however, any measure that does not produce scores by some random process is by that definition generalizable to some other performance and, hence, to that extent may be said to be valid. What a given measure is valid for, that is, generalizable to, however, is a matter of discovery as much as of plan. All instruments used in clinical assessment should be subjected to comprehensive and continuing investigation in order to determine the sources of variance in scores. An instrument that has good generalizability over time and across raters may turn out to be, among other things, a very good measure of some response style or other bias. The MMPI includes a number of “validity” scales designed to assess various biases in performance on it, and it has been subjected to many investigations of bias. The same cannot be said of some other widely used clinical assessment instruments and procedures. To take the most notable example, of the more than 1000 articles on the Rorschach that are in the current PsychInfo database, only a handful, about 1%, appear to deal with issues of response bias, and virtually all of those are on malingering and most of them are unpublished dissertations.

4.01.3.3 Item Response Theory

Although Item Response Theory (IRT) is a potentially powerful tool for the development and study of measures of many kinds, its use to date has not been extensive beyond the area of

ability testing. The origins of IRT go back at least to the early 1950s and the publication of Lord's (1952) monograph, *A theory of test scores*, but it has had little impact on measurement outside the arena of ability testing (Meier, 1994). Certainly it has had almost no impact on clinical assessment. The current PsychInfo database includes only two references to IRT in relation to the MMPI and only one to the Rorschach, and the latter one, now 10 years old, is an entirely speculative mention of a potential application of IRT (Samejima, 1988).

IRT, perhaps to some extent narrowly imagined to be relevant only to test construction, can be of great value in exploring the nature of measures and improving their interpretation. For example, IRT can be useful in understanding just when scores may be interpreted as unidimensional and then in determining the size of gaps in underlying traits represented by adjacent scores. An example could be the interpretation of Whole responses on the Rorschach. Is the W score a unidimensional score, and, if so, is each increment in that score to be interpreted as an equal increment? Some cards are almost certainly more difficult stimuli to which to produce a W response, and IRT could calibrate that aspect of the cards. IRT would be even more easily used for standard paper-and-pencil inventory measures, but the total number of applications to date is small, and one can only conclude that clinical assessment is being short-changed in its development.

4.01.3.4 Scores on Tests

Lord's (1952) monograph was aimed at tests with identifiable underlying dimensions such as ability. Clinical assessment appears never to have had any theory of scores on instruments included under that rubric. That is, there seems never to have been proposed or adapted any unifying theory about how test scores on clinical instruments come about. Rather there seems to have been a passive, but not at all systematic, adoption of general test theory, that is, the idea that test scores are in some manner generated by responses representing some underlying trait. That casual approach cannot forward the development of the field.

Fiske (1971) has come about as close as anyone to formulating a theory of test scores for clinical assessment, although his ideas pertain more to how such tests are scored than to how they come about, and his presentation was directed toward personality measurement rather than clinical assessment. He suggested several models for scoring test, or otherwise observed, responses. The simplest model is what we may call the cumulative frequency model,

which simply increments the score by 1 for every observed response. This is the model that underlies many Rorschach indices. It assumes that every response is equivalent to every other one, and it ignores the total number of opportunities for observation. Thus, each Rorschach W response counts as 1 for that index, and the index is not adjusted to take account of the total number of responses. A second model is the relative frequency model, which forms an index by dividing the number of observed critical responses by some indicator of opportunities to form a rate of responding, for example, as would be accomplished by counting W responses and dividing by the total number of responses or by counting W responses only for the first response to each card. Most paper-and-pencil inventories are scored implicitly in that way, that is, they count the number of critical responses in relation to the total number possible.

A long story must be made short here, but Fiske describes other models, and still more are possible. One may weight responses according to the inverse of their frequency in a population on the grounds that common responses should count for less than rare responses. Or one may weight responses according to the judgments of experts. One can assign the average weight across a set of responses, a common practice, but one can also assign as the score the weight of the most extreme response, for example, as runners are often rated on the basis of their fastest time for any given distance. Pathology is often scored in that way, for example, a pathognomic response may outweigh many mundane, ordinary responses.

The point is that clinical assessment instruments and procedures only infrequently have any explicit basis in a theory of responses. For the most part, scores appear to be derived in some standard way without much thought having been given to the process. It is not clear how much improvement in measures might be achieved by more attention to the development of a theory of scores, but it surely could not hurt to do so.

4.01.3.5 Calibration of Measures

A critical limitation on the utility of psychological measures of any kind, but certainly in their clinical application, is the fact that the measures do not produce scores in any directly interpretable metric. We refer to this as the *calibration* problem (Sechrest, McKnight, & McKnight, 1996). The fact is that we have only a very general knowledge of how test scores may be related to any behavior of real interest. We may know in general that a score of 70, let us

say, on an MMPI scale is “high,” but we do not know very well what might be expected in the behavior of a person with such a score. We would know even less about what difference it might make if the score were reduced to 60 or increased to 80 except that in one case we might expect some diminution in problems and in the other some increase. In part the lack of calibration of measures in clinical psychology stems from lack of any specific interest and diligence in accomplishing the task. Clinical psychology has been satisfied with “loose calibration,” and that stems in part, as we will assert later, from adoption of the uninformative model of significance testing as a standard for validation of measures.

4.01.4 WHY HAVE WE MADE SO LITTLE PROGRESS?

It is difficult to be persuaded that progress in assessment in clinical psychology has been substantial in the past 75 years, that is, since the introduction of the Rorschach. Several arguments may be adduced in support of that statement, even though we recognize that it will be met with protests. We will summarize what we think are telling arguments in terms of theory, formats, and validities of tests.

First, we do not discern any particular improvements in theories of clinical testing and assessments over the past 75 years. The Rorschach, and the subsequent formulation of the projective hypothesis, may be regarded as having been to some extent innovations; they are virtually the last ones in the modern history of assessment. As noted, clinical assessment lags well behind the field in terms of any theory of either the stimuli or responses with which it deals, let alone the connections between them. No theory of assessment exists that would guide selection of stimuli to be presented to subjects, and certainly none pertains to the specific format of the stimuli nor to the nature of the responses required. Just to point to two simple examples of the deficiency in understanding of response options, we note that there is no theory to suggest whether in the case of a projective test responses should be followed by any sort of inquiry about their origins, and there is no theory to suggest in the case of self-report inventories whether items should be formulated so as to produce endorsements of the “this is true of me” nature or so as to produce descriptions such as “this is what I do.”

Given the lack of any gains in theory about the assessment enterprise, it is not surprising that there have also not been any changes in test formats since the introduction of the Rorschach.

Projective tests based on the same simple (and inadequate) hypothesis are still being devised, but not one has proven itself in any way better than anything that has come before. Item writers may be a bit more sophisticated than those in the days of the Bernreuter, but items are still constructed in the same way, and response formats are the same as ever, “agree–disagree,” “true–false,” and so on.

Even worse, however, is the fact that absolutely no evidence exists to suggest that there have been any mean gains in the validities of tests over the past 75 years. Even for tests of intellectual functioning, typical correlations with any external criterion appear to average around 0.40, and for clinical and personality tests the typical correlations are still in the range of 0.30, the so-called “personality coefficient.” This latter point, that validities have remained constant, may, of course, be related to the lack of development of theory and to the fact that the same test formats are still in place.

Perhaps some psychologists may take exception to the foregoing and cite considerable advances. Such claims are made for the Exner (1986) improvements on the Rorschach, known as the “comprehensive system,” and for the MMPI-2, but although both claims are superficially true, there is absolutely no evidence for either claim from the standpoint of validity of either test. The Exner comprehensive system seems to have “cleaned up” some aspects of Rorschach scoring, but the improvements are marginal, for example, it is not as if inter-rater reliability increased from 0.0 to 0.8, and no improvements in validity have been established. Even the improvements in scoring have been demonstrated for only a portion of the many indexes. The MMPI-2 was only a cosmetic improvement over the original, for example, getting rid of some politically incorrect items, and no increase in the validity of any score or index seems to have been demonstrated, nor is any likely.

An additional element in the lack of evident “progress” in the validity of test scores may be lack of reliability (and validity!) of people being predicted. (One wise observer suggested that we would not really like it at all if behavior were 90% predictable! Especially our own.) We may just have reached the limits of our ability to predict what is going to happen with and to people, especially with our simple-minded and limited assessment efforts. As long as we limit our assessment efforts to the dispositions of the individuals who are clients and ignore their social milieus, their real environmental circumstances, their genetic possibilities, and so on, we may not be able to get beyond correlations of 0.3 or 0.4.

The main “advance” in assessment over the past 75 years is not that we do anything really better but that we do it much more widely. We have many more scales than existed in the past, and we can at least assess more things than ever before, even if we can do that assessment only, at best, passably well.

Woodworth (1937/1992) wrote in his article on the future of clinical psychology that, “There can be no doubt that it will advance, and in its advance throw into the discard much guesswork and half-knowledge that now finds baleful application in the treatment of children, adolescents and adults” (p. 16). It appears to us that the opposite has occurred. Not only have we failed to discard guesswork and half-knowledge, that is, tests and treatments with years of research indicating little effect or utility, we have continued to generate procedures based on the same flawed assumptions with the misguided notion that if we just make a bit of a change here and there, we will finally get it right. Projective assessments that tell us, for example, that a patient is psychotic are of little value. Psychologists have more reliable and less expensive ways of determining this. More direct methods have higher validity in the majority of cases. The widespread use of these procedures at high actual and opportunity cost is not justified by the occasional addition of information. It is not possible to know ahead of time which individuals might give more information via an indirect method, and most of the time it is not even possible to know afterwards whether indirectly obtained “information” is correct unless the information has also been obtained in some other way, that is, asking the person, asking a relative, or doing a structured interview. It is unlikely that projective test responses will alter clinical intervention in most cases, nor should it.

Is it fair to say that clinical psychology has no standards (see Sechrest, 1992)? Clinical psychology gives the appearance of standards with accreditation of programs, internships, licensure, ethical standards, and so forth. It is our observation, however, that there is little to no monitoring of the purported standards. For example, in reviewing recent literature as background to this chapter, we found articles published in peer-reviewed journals using projective tests as outcome measures for treatment. The APA ethical code of conduct states that psychologists “. . . use psychological assessment . . . for purposes that are appropriate in light of the research on or evidence of the . . . proper application of the techniques.” The APA document, *Standards for educational and psychological testing*, states:

... Validity however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (APA, 1985, p. 9)

Further, the section titled, *Professional standards for test use* (APA, 1985, p. 42, Standard 6.3) states:

When a test is to be used for a purpose for which it has not been previously validated, or for which there is no supported claim for validity, the user is responsible for providing evidence of validity.

No body of research exists to support the validity of any projective instrument as the sole outcome measure for treatment—or as the sole measure of anything. So not only do questionable practices go unchecked, they can result in publication.

4.01.4.1 The Absence of the Autopsy

Medicine has always been disciplined by the regular occurrence of the autopsy. A physician makes a diagnosis and treats a patient, and if the patient dies, an autopsy will be done, and the physician will receive feedback on the correctness of his or her diagnosis. If the diagnosis were wrong, the physician would to some extent be called to account for that error; at least the error would be known, and the physician could not simply shrug it off. We know that the foregoing is idealized, that autopsies are not done in more than a fraction of cases, but the model makes our point. Physicians make predictions, and they get feedback, often quickly, on the correctness of those predictions. Surgeons send tissue to be biopsied by pathologists who are disinterested; internists make diagnoses based on various signs and symptoms and then order laboratory procedures that will inform them about the correctness of their diagnosis; family practitioners make diagnoses and prescribe treatment, which, if it does not work, they are virtually certain to hear about.

Clinical psychology has no counterpart to the autopsy, no systematic provision for checking on the correctness of a conclusion and then providing feedback to the clinician. Without some form of systematic checking and feedback, it is difficult to see how improvement in either instruments or clinicians' use of them could be regularly and incrementally improved. Psychologist clinicians have been allowed the slack involved in making unbounded predictions and then not getting any sort of feedback on the

potential accuracy of even those loose predictions. We are not sure how much improvement in clinical assessment might be possible even with exact and fairly immediate feedback, but we are reasonably sure that very little improvement can occur without it.

4.01.5 FATEFUL EVENTS CONTRIBUTING TO THE HISTORY OF CLINICAL ASSESSMENT

The history of assessment in clinical psychology is somewhat like the story of the evolution of an organism in that at critical junctures, when the development of assessment might well have gone one way, it went another. We want to review here several points that we consider to be critical in the way clinical assessment developed within the broader field of psychology.

4.01.5.1 The Invention of the Significance Test

The advent of hypothesis testing in psychology had fateful consequences for the development of clinical assessment, as well as for the rest of psychology (Gigerenzer, 1993). Hypothesis testing encouraged a focus on the question whether any predictions or other consequences of assessment were "better than chance," a distinctly loose and undemanding criterion of "validity" of assessment. The typical validity study for a clinical instrument would identify two groups that would be expected to differ in some "score" derived from the instrument and then ask the question whether the two groups did in fact (i.e., to a statistically significant degree) differ in that score. It scarcely mattered by how much they differed or in what specific way, for example, an overall mean difference vs. a difference in proportions of individuals scoring beyond some extreme or otherwise critical value. The existence of any "significant" difference was enough to justify triumphant claims of validity.

4.01.5.2 Ignoring Decision Making

One juncture had to do with bifurcation of the development of clinical psychology from other streams of assessment development. Specifically, intellectual assessment and assessment of various capacities and propensities relevant to performance in work settings veered in the direction of assessment for decision-making (although not terribly sharply nor completely), while assessment in clinical psychology went in the direction of assessment for enlightenment. What eventually happened is that clinical psychology failed to adopt any rigorous

criterion of correctness of decisions made on the basis of assessed performance, but adopted instead a conception of assessments as generally informative or "correct."

Simply to make the alternative clear, the examples provided by medical assessment are instructive. The model followed in psychology would have resulted in medical research of some such nature as showing that two groups that "should" have differed in blood pressure, for example, persons having just engaged in vigorous exercise vs. persons having just experienced a rest period, differed significantly in blood pressure readings obtained by a sphygmomanometer. Never mind by how much they differed or what the overlap between the groups. The very existence of a "significant" difference would have been taken as evidence for the "validity" of the sphygmomanometer.

Instead, however, medicine focused more sharply on the accuracy of decisions made on the basis of assessment procedures. The aspect of biomedical assessment that most clearly distinguishes it from clinical psychological assessment is its concern for sensitivity and specificity of measures (instruments) (Kraemer, 1992). Kraemer's book, *Evaluating medical tests: Objective and quantitative guidelines*, has not even a close counterpart in psychology, which is, itself, revealing. These two characteristics of measures are radically different from the concepts of validity used in psychology, although "criterion validity" (now largely abandoned) would seem to require such concepts.

Sensitivity refers to the proportion of cases having a critical characteristic that are identified by the test. For example, if a test were devised to select persons likely to benefit from some form of therapy, sensitivity would refer to the proportion of cases that would actually benefit which would be identified correctly by the test. These cases would be referred to as "true positives." Any cases that would benefit from the treatment but that could not be identified by the test would be "false-negatives" in this example. Conversely, a good test should have high specificity, which would be avoiding "false-positives," or incorrectly identifying as good candidates for therapy persons who would not actually benefit. The "true negative" group would be those persons who would not benefit from treatment, and a good test should correctly identify a large proportion of them.

As Kraemer (1992) points out, sensitivity and specificity as test requirements are nearly always in opposition to each other, and are reciprocal. Maximizing one requirement reduces the other. Perfect sensitivity can be attained by, in our example, a test that identifies every case as suitable for therapy; no amenable cases are

missed. Unfortunately, that maneuver would also maximize the number of false-positives, that is, many cases would be identified as suitable for therapy who, in fact, were not. Obviously, the specificity of the test could be maximized by declaring all cases as unsuitable for therapy, thus ensuring that the number of false-positives would be zero—while at the same time ensuring that the number of false-negatives would be maximal, and no one would be treated.

We go into these issues in some detail in order to make clear how very different such thinking is from usual practices in clinical psychological assessment. The requirements for Receiver Operating Curves (ROC), which is the way issues of sensitivity and specificity of measures are often labeled and portrayed, are stringent. They are not satisfied by simple demonstrations that measures, for example, suitability for treatment, are "significantly related to" other measures of interest, for example, response to treatment. The development of ROC statistics almost always occurs in the context of the use of tests for decision-making: treat—not treat, hire—not hire, do further tests—no further tests. Those kinds of uses of tests in clinical psychological assessment appear to be rare.

Issues of sensitivity-specificity require the existence of some reasonably well-defined criterion, for example, the definition of what is meant by favorable response to treatment and a way of measuring it. In biomedical research, ROC statistics are often developed in the context of a "gold standard," a definitive criterion. For example, an X ray might serve as a gold standard for a clinical judgment about the existence of a fracture, or a pathologist's report on a cytological analysis might serve as a gold standard for a screening test designed to detect cancer. Clinical psychology has never had anything like a gold standard against which its various tests might have been validated. Psychiatric diagnosis has sometimes been of interest as a criterion, and tests of different types have been examined to determine the extent to which they produce a conclusion in agreement with diagnosis (e.g., Somoza, Steer, Beck, & Clark, 1994), but in that case the gold standard is suspect, and it is by no means clear that disagreement means that the test is wrong.

The result is that for virtually no psychological instrument is it possible to produce a useful quantitative estimate of its accuracy. Tests and other assessment devices in clinical psychology have been used for the most part to produce general enlightenment about a target of interest rather than to make a specific prediction of some outcome. People who have been tested are described as "high in anxiety," "clinically

depressed,” or “of average intelligence.” Statements of that sort, which we have referred to previously as unbounded predictions, are possibly enlightening about the nature of a person’s functioning or about the general range within which problems fall, but they are not specific predictions, and are difficult to refute.

4.01.5.3 Seizing on Construct Validity

In 1955, Cronbach and Meehl published what is arguably the most influential article in the field of measurement: *Construct validity in psychological tests* (Cronbach & Meehl, 1955). This is the same year as the publication of *Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores* (Meehl & Rosen, 1955). It is safe to say that no two more important articles about measurement were ever published in the same year. The propositions set forth by Cronbach and Meehl about the validity of tests were provocative and rich with implications and opportunities. In particular, the idea of construct validity required that measures be incorporated into elaborated theoretical structure, which was labeled the “nomological net.” Unfortunately, the fairly daunting requirements for embedding measures in theory were mostly ignored in clinical assessment (the same could probably be said about most other areas of psychology, but it is not our place here to say so), and the idea of construct validity was trivialized.

The trivialization of construct validity reflects in part the fact that no standards for construct validity exist (and probably none can be written) and the general failure to distinguish between necessary and sufficient conditions for the inference of construct validity. In their presentation of construct validity, Cronbach and Meehl did not specify any particular criteria for sufficiency of evidence, and it would be difficult to do so. Construct validity exists when everything fits together, but trying to specify the number and nature of the specific pieces of evidence would be difficult and, perhaps, antithetical to the idea itself. It is also not possible to quantify level or degree of construct validity other than in a very rough way and such quantifications are, in our experience, rare. It is difficult to think of an instance of a measure described as having “moderate or “low” construct validity, although “high” construct validity is often implied.

It is possible to imagine what some of the necessary conditions for construct validity might be, one notable requirement being convergent validity (Campbell & Fiske, 1959). In some manner that we have not tried to trace, conditions necessary for construct validity came

to be viewed as sufficient. Thus, for example, construct validity usually requires that one measure of a construct correlates with another. Such a correlation is not, however, a sufficient condition for construct validity, but, nonetheless, a simple zero-order correlation between two tests is often cited as “evidence” for the construct validity of one measure or the other. Even worse, under the pernicious influence of the significance testing paradigm, any statistically significant correlation may be taken as evidence of “good construct validity.” Or, for another example, construct validity usually requires a particular factor structure for a measure, but the verification of the required factor structure is not sufficient evidence for construct validity of the measure involved. The fact that a construct is conceived as unidimensional does not mean that a measure alleged to represent the construct does so simply because it appears to form a single factor.

The net result of the dependence on significance testing and the poor implementation of the ideas represented by construct validity has been that the standards of evidence for the validity of psychological measures has been distressingly low.

4.01.5.4 Adoption of the Projective Hypothesis

The projective hypothesis (Frank, 1939) is a general proposition stating that whatever an individual does when exposed to an ambiguous stimulus will reveal important aspects of his or her personality. Further, the projective hypothesis suggests that indirect responses, that is, those to ambiguous stimuli, are more valid than direct responses, that is, those to interviews or questionnaires. There is little doubt that indirect responses reveal something about people, although whether that which is revealed is, in fact, important is more doubtful. Moreover, what one eats, wears, listens to, reads, and so on are rightly considered to reveal something about that individual. While the general proposition about responses to ambiguous stimuli appears quite reasonable, the use of such stimuli in the form of projective tests has proven problematic and of limited utility.

The course of development of clinical assessment might have been different and more useful had it been realized that projection was the wrong term for the link between ambiguous stimuli and personality. A better term would have been the “expressive hypothesis,” the notion that an individual’s personality may be manifest (expressed) in response to a wide range of stimuli, including ambiguous stimuli. Personality style might have come to be of greater concern, and unconscious determinants

of behavior, implied by projection, might have received less emphasis.

In any case, when clinical psychology adopted the projective hypothesis and bought wholesale into the idea of unconscious determinants of behavior, that set the field on a course that has been minimally productive but that still affects an extraordinarily wide range of clinical activities. Observable behaviors have been downplayed and objective measures treated with disdain or dismissed altogether. The idea of peering into the unconscious appealed both to psychological voyeurs and to those bent on achieving the glamour attributed to the psychoanalyst.

Research on projective stimuli indicates that highly structured stimuli which limit the dispositions tapped increase the reliability of such tests (e.g., Kagan, 1959). In achieving acceptable reliability, the nature of the test is altered in such a way that the stimulus is less ambiguous and the likelihood of an individual "projecting" some aspect of their personality in an unusual way becomes reduced. Thus, the dependability of responses to projective techniques probably depends to an important degree on sacrificing their projective nature. In part, projective tests seem to have failed to add to assessment information because most of the variance in responses to projective stimuli is accounted for by the stimuli themselves. For example, "popular" responses on the Rorschach are popular because the stimulus is the strongest determinant of the response (Murstein, 1963).

Thorndike (Thorndike & Hagen, 1955, p. 418), in describing the state of affairs with projective tests some 40 years ago, stated:

A great many of the procedures have received very little by way of rigorous and critical test and are supported only by the faith and enthusiasm of their backers. In those few cases, most notable that of the Rorschach, where a good deal of critical work has been done, results are varied and there is much inconsistency in the research picture. Modest reliability is usually found, but consistent evidence of validity is harder to come by.

The picture has not changed substantially in the ensuing 40 years and we doubt that it is likely to change much in the next 40. As Adcock (1965, cited in Anastasi, 1988) noted, "There are still enthusiastic clinicians and doubting statisticians." As noted previously (Sechrest, 1963, 1968), these expensive and time-consuming projective procedures add little if anything to the information gained by other methods and their abandonment by clinical psychology would not be a great loss. Despite lack of incremental validity after decades of research,

not only do tests such as the Rorschach and TAT continue to be used, but new projective tests continue to be developed. That could be considered a pseudoscientific enterprise that, at best, yields procedures telling clinical psychologists what they at least should already know or have obtained in some other manner, and that, at worst, wastes time and money and further damages the credibility of clinical psychology.

4.01.5.5 The Invention of the Objective Test

At one time we had rather supposed without thinking about it too much that objective tests had always been around in some form or other. Samelson (1987), however, has shown that at least the multiple-choice test was invented in the early part of the twentieth century, and it seems likely that the true-false test had been devised not too long before then. The objective test revolutionized education in ways that Samelson makes clear, and it was not long before that form of testing infiltrated into psychology. Bernreuter (1933) is given credit for devising the first multiphasic (multidimensional) personality inventory—only 10 years after the introduction of the Rorschach into psychology.

Since 1933, objective tests have flourished. In fact, they are now much more widely used than projective tests and are addressed toward almost every imaginable problem and aspect of human behavior. The Minnesota Multiphasic Personality Inventory (1945) was the truly landmark event in the course of development of paper-and-pencil instruments for assessing clinical aspects of psychological functioning. "Paper-and-pencil" is often used synonymously with "objective" in relation to personality. From that time on, other measures flourished, of recent in great profusion.

Paper-and-pencil tests freed clinicians from the drudgery of test administration, and in that way they also made testing relatively inexpensive as a clinical enterprise. They also made tests readily available to psychologists not specifically trained on them, including psychologists at subdoctoral levels. Paper-and-pencil measures also seemed so easy to administer, score, and interpret. As we have noted previously, the ease of creation of new measures had very substantial effects on the field, including clinical assessment.

4.01.5.6 Disinterest in Basic Psychological Processes

Somewhere along the way in its development, clinical assessment became detached from the mainstream of psychology and, therefore, from

the many developments in basic psychological theory and knowledge. The Rorschach was conceived not as a test of personality *per se* but in part as an instrument for studying perception and Rorschach referred to it as his "experiment" (Hunt, 1956). Unfortunately, the connections of the Rorschach to perception and related mental processes were lost, and clinical psychology became preoccupied not with explaining how Rorschach responses come to be made but in explaining how Rorschach responses reflect back on a narrow range of potential determinants: the personality characteristics of respondents, and primarily their pathological characteristics at that.

It is testimony to the stasis of clinical assessment that three-quarters of a century after the introduction of the Rorschach, a period of time marked by stunning (relatively) advances in understanding of such basic psychological processes as perception, cognition, learning, and motivation and by equivalent or even greater advances in understanding of the biological structures and processes that underlie human behavior, the Rorschach continues, virtually unchanged, to be the favorite instrument for clinical assessment. The Exner System, although a revision of the scoring system, in no way reflects any basic changes in our advancement of understanding of the psychological knowledge base in which the Rorschach is, or should be, embedded. Take, just for one instance, the great increase of interest in and understanding of "priming" effects in cognition; those effects would clearly be relevant to the understanding of Rorschach responses, but there is no indication at all of any awareness on the part of those who write about the Rorschach that any such effect even exists. It was known a good many years ago that Rorschach responses could be affected by the context of their administration (Sechrest, 1968), but without any notable effect on their use in assessment.

Nor do any other psychological instruments show any particular evidence of any relationship to the rest of the field of psychology. Clinical assessment could have benefited greatly from a close and sensitive connection to basic research in psychology. Such a connection might have fostered interest in clinical assessment in the development of instruments for the assessment of basic psychological processes.

Clinical psychology has—is afflicted with, we might say—an extraordinary number of different tests, instruments, procedures, and so on. It is instructive to consider the nature of all these tests; they are quite diverse. (We use the term "test" in a somewhat generic way to refer to the wide range of mechanisms by which psychol-

ogists carry out assessments.) Whether the great diversity is a curse or a blessing depends on one's point of view. We think that a useful perspective is provided by contrasting psychological measures with those typically used in medicine, although, obviously, a great many differences exist between the two enterprises. Succinctly, however, we can say that most medical tests are very narrow in their intent, and they are devised to tap basic states or processes. A screening test for tuberculosis, for example, involves subcutaneous injection of tuberculin which, in an infected person, causes an inflammation at the point of injection. The occurrence of the inflammation then leads to further narrowly focused tests. The inflammation is not tuberculosis but a sign of its potential existence. A creatinine clearance test is a test of renal function based on the rate of clearance of ingested creatinine from the blood. A creatinine clearance test can indicate abnormal renal functioning, but it is a measure of a fundamental physiological process, not a state, a problem, a disease, or anything of that sort. A physician who is faced with the task of diagnosing some disease process involving renal malfunction will use a variety of tests, not necessarily specified by a protocol (battery) to build an information base that will ultimately lead to a diagnosis.

By contrast, psychological assessment is, by and large, not based on measurement of basic psychological processes, with few exceptions. Memory is one function that is of interest to neuropsychologists, and occasionally to others, and instruments to measure memory functions do exist. Memory can be measured independently of any other functions and without regard to any specific causes of deficiencies. Reaction time is another basic psychological process. It is currently used by cognitive psychologists as a proxy for mental processing time, and since the 1970s, interest in reaction time as a marker for intelligence has grown and become an active research area.

For the most part, however, clinical assessment has not been based on tests of basic psychological functions, although the Wechsler intelligence scales might be regarded as an exception to that assertion. A very large number of psychological instruments and procedures are aimed at assessing syndromes or diagnostic conditions, whole complexes of problems. Scales for assessing attention deficit disorder (ADD), suicide probability, or premenstrual syndrome (PMS) are instances. Those instruments are the equivalent of a medical "Test for Diabetes," which does not exist. The Conners' Rating Scales (teachers) for ADD, for example, has subscales for Conduct Problem, Hyperactivity, Emotional Overindulgent, Asocial,

Anxious-Passive, and Daydream-Attendance. Several of the very same problems might well be represented on other instruments for entirely different disorders. But if they were, they would involve a different set of items, perhaps with a slightly different twist, to be integrated in a different way. Psychology has no standard ways of assessing even such fundamental dispositions as “asocial.”

One advantage of the medical way of doing things is that tests like creatinine clearance have been used on millions of persons, are highly standardized, have extremely well-established norms, and so on. Another set of ADD scales, the Brown, assesses “ability to activate and organize work tasks.” That sounds like an important characteristic of children, so important that one might think it would be widely used and useful. Probably, however, it appears only on the Brown ADD Scales, and it is probably little understood otherwise.

Clinical assessment has also not had the benefit of careful study from the standpoint of basic psychological processes that affect the clinician and his or her use and interpretation of psychological tests. Achenbach (1985), to cite a useful perspective, discusses clinical assessment in relation to the common sources of error in human judgment. Achenbach refers to such problems as illusory correlation, inability to assess covariation, and the representativeness and availability heuristics and confirmatory bias described by Kahneman, Slovic, and Tversky (1982). Consideration of these sources of *human*, that is, general, error in judgment would be more likely if clinical assessment were more attuned to and integrated into the mainstream developments of psychology.

We do not suppose that clinical assessment should be limited to basic psychological processes; there may well be a need for syndrome-oriented or condition-oriented instruments. Without any doubt, however, clinical assessment would be on a much firmer footing if from the beginning psychologists had tried to define and measure well a set of fundamental psychological processes that could be tapped by clinicians faced with diagnostic or planning problems.

Unfortunately, measurement has never been taken seriously in psychology, and it is still lightly regarded. One powerful indicator of the casual way in which measurement problems are met in clinical assessment is the emphasis placed on brevity of measures. “. . . entire exam can be completed. . . in just 20 to 30 minutes” (for head injury), “completed in just 15–20 minutes” (childhood depression), “39 items” (to measure six factors involved in ADD) are just a few of the notations concerning tests that are brought to

the attention of clinician-assessors by advertisers. It would be astonishing to think of a medical test advertised as “diagnoses brain tumors in only 15 minutes,” or “complete diabetes workup in only 30 minutes.” An MRI examination for a patient may take up to several hours from start to finish, and no one suggests a “short form” of one. Is it imaginable that one could get more than the crudest notion of childhood depression in 15–20 minutes?

4.01.6 MISSED SIGNALS

At various times in the development of clinical psychology, opportunities existed to guide, or even redirect, assessment activities in one way or another. Clinical psychology might very well have taken quite a different direction than it has (Sechrest, 1992). Unfortunately, in our view, a substantial number of critical “signals to the field were missed, and entailed in missing them was failure to redirect the field in what would have been highly constructive ways.

4.01.6.1 The Scientist–Practitioner Model

We do not have the space to go into the intricacies of the scientist–practitioner model of training and practice, but it appears to be an idea whose time has come and gone. Suffice it to say here that full adoption of the model would not have required every clinical practitioner to be a researcher, but it would have fostered the idea that to some extent every practitioner is responsible for the scientific integrity of his or her own practice, including the validity of assessment procedures. The scientist–practitioner model might have helped clinical psychologists to be involved in research, even if only as contributors rather than as independent investigators.

That involvement could have been of vital importance to the field. The development of psychological procedures will never be supported commercially to any appreciable extent, and if they are to be adequately developed, it will have to be with the voluntary—and enthusiastic—participation of large numbers of practitioners who will have to contribute data, be involved in the identification of problems, and so on. That participation would have been far more likely had clinical psychology stuck to its original views of itself (Sechrest, 1992).

4.01.6.2 Construct Validity

We have already discussed construct validity at some length, and we have explained our view

that the idea has been trivialized, in essence abandoned. That is another lost opportunity, because the power of the original formulation by Cronbach and Meehl (1955) was great. Had their work been better understood and honestly adopted, clinical psychology would by this time almost certainly have had a set of well-understood and dependable measures and procedures. The number and variety of such measures would have been far less than exists now, and the dependability of them would have been circumscribed, but surely it would have been better to have good than simply many measures.

4.01.6.3 Assumptions Underlying Assessment Procedures

In 1952, Lindzey published a systematic analysis of assumptions underlying the use of projective techniques (Lindzey, 1952). His paper was a remarkable achievement, or would have been had anyone paid any attention to it. The Lindzey paper could have served as a model and stimulus for further formulations leading to a theory, comprehensive and integrated, of performance on clinical instruments. A brief listing of several of the assumptions must suffice to illustrate what he was up to:

IV. The particular response alternatives emitted are determined not only by characteristic response tendencies (enduring dispositions) but also by intervening defenses and his cognitive style.

XI. The subject's characteristic response tendencies are sometimes reflected indirectly or symbolically in the response alternatives selected or created in the test situation.

XIII. Those responses that are elicited or produced under a variety of different stimulus conditions are particularly likely to mirror important aspects of the subject.

XV. Responses that deviate from those typically made by other subjects to this situation are more likely to reveal important characteristics of the subject than modal responses which are more like those made by most other subjects.

These and other assumptions listed by Lindzey could have provided a template for systematic development of both theory and programs of research aimed at supporting the empirical base for projective—and other—testing. Assumption XI, for example, would lead rather naturally to the development of explicit theory, buttressed by empirical data, which would indicate just when responses probably should and should not be interpreted as symbolic.

Unfortunately, Lindzey's paper appears to have been only infrequently cited and to have been substantially ignored by those who were engaged in turning out all those projective tests, inventories, scales, and so on. At this point we know virtually nothing more about the performance of persons on clinical instruments than was known by Lindzey in 1952. Perhaps even less.

4.01.6.4 Antecedent Probabilities

In 1955 Meehl and Rosen published an exceptional article on antecedent probabilities and the problem of base rates. The article was, perhaps, a bit mathematical for clinical psychology, but it was not really difficult to understand, and its implications were clear. Whenever one is trying to predict (or diagnose) a characteristic that is quite unevenly distributed in a population, the difficulty in beating the accuracy of the simple base rates is formidable, sometimes awesomely so. For example, even in a population considered at high risk for suicide, only a very few persons will actually commit suicide. Therefore, unless a predictive measure is extremely precise, the attempt to identify those persons who will commit suicide will identify as suicidal a relatively large number of "false-positives," that is, if one wishes to be sure not to miss any truly suicidal people, one will include in the "predicted suicide" group a substantial number of people not so destined. That problem is a serious to severe limitation when the cost of missing a true-positive is high, but so, relatively, is the cost of having to deal with a false-positive.

More attention to the difficulties described by Meehl and Rosen (1955) would have moved psychological assessment in the direction taken by medicine, that is, the use of ROCs. Although ROCs do not make the problem go away, they keep it in the forefront of attention and require that those involved, whether researchers or clinicians, deal with it. That signal was missed in clinical psychology, and it is scarcely mentioned in the field today. Many indications exist that a large proportion of clinical psychologists are quite unaware that the problem even exists, let alone that they have an understanding of it.

4.01.6.5 Need for Integration of Information

Many trends over the years converge on the conclusion that psychology will make substantial progress only to the extent that it is able to integrate its theories and knowledge base with those developing in other fields. We can address this issue only on the basis of personal experience; we can find no evidence for our

view. Our belief is that clinical assessment in psychology rarely results in a report in which information related to a subject's genetic disposition, family structure, social environment, and so on are integrated in a systematic and effective way.

For example, we have seen many reports on patients evaluated for alcoholism without any attention, let alone systematic attention, to a potential genetic basis for their difficulty. At most a report might include a note to the effect that the patient has one or more relatives with similar problems. Never was any attempt made to construct a genealogy that would include other conditions likely to exist in the families of alcoholics. The same may be said for depressed patients. It might be objected that the responsibilities of the psychologist do not extend into such realms as genetics and family and social structure, but surely that is not true if the psychologist aspires to be more than a sheer technician, for example, serving the same function as a laboratory technician who provides a number for the creatinine clearance rate and leaves it to someone else, "the doctor," to put it all together.

That integration of psychological and other information is of great importance has been implicitly known for a very long time. That knowledge has simply never penetrated training programs and clinical practice. That missed opportunity is to the detriment of the field.

4.01.6.6 Method Variance

The explicit formulation of the concept of method variance was an important development in the history of assessment, but one whose import was missed or largely ignored. The concept is quite simple: to some extent, the value obtained for the measurement of any variable depends in part on the characteristics of the method used to obtain the estimate. (A key idea is the understanding that any specific value is, in fact, an estimate.) The first explicit formulation of the idea of method variance was the seminal Campbell and Fiske paper on the "multitrait-multimethod matrix" (Campbell & Fiske, 1959). (That paper also introduced the very important concepts of "convergent" and "discriminant" validity, now widely employed but, unfortunately, not always very well understood.) There had been precursors of the idea of method variance. In fact, much of the interest in projective techniques stemmed from the idea that they would reveal aspects of personality that would not be discernible from, for example, self-report measures. The MMPI, first published in 1943 (Hathaway & McKinley),

included "validity" scales that were meant to detect, and, in the case of the K-scale, even correct for, method effects such as lying, random responding, faking, and so on. By 1960 or so, Jackson and Messick had begun to publish their work on response styles in objective tests, including the MMPI (e.g., Jackson & Messick, 1962). At about the same time, Berg (1961) was describing the "deviant response tendency," which was the hypothesis that systematic variance in test scores could be attributed to general tendencies on the part of some respondents to respond in deviant ways. Nonetheless, it was the Campbell and Fiske (1959) paper that brought the idea of method variance to the attention of the field.

Unfortunately, the cautions expressed by Campbell and Fiske, as well as by others working on response styles and other method effects, appear to have had little effect on developments in clinical assessment. For the most part, the problems raised by method effects and response styles appear to have been pretty much ignored in the literature on clinical assessment. A search of a current electronic database in psychology turned up, for example, only one article over the past 30 years or so linking the Rorschach to any discussion of method effects (Meyer, 1996). When one considers the hundreds of articles having to do with the Rorschach that were published during that period of time, the conclusion that method effects have not got through to the attention of the clinical assessment community is unavoidable. The consequence almost surely is that clinical assessments are not being corrected, at least not in any systematic way, for method effects and response biases.

4.01.6.7 Multiple Measures

At least a partial response to the problem of method effects in assessment is the use of multiple measures, particularly measures that do not appear to share sources of probable error or bias. That recommendation was explicit in Campbell and Fiske (1959), and it was echoed and elaborated upon in 1966 (Webb et al., 1966), and again in 1981 (Webb et al., 1981). Moreover, Webb and his colleagues warned specifically against the very heavy reliance on self-report measures in psychology (and other social sciences). That warning, too, appears to have made very little difference in practice. Examination of catalogs of instruments meant to be used in clinical assessment will show that a very large proportion of them depend upon self-reports of individual subjects about their own dispositions, and measures that do not rely

directly on self-reports nonetheless do nearly all rely solely on the verbal responses of subjects. Aside from rating scales to be used with parents, teachers, or other observers of behavior, characteristics of interest such as personality and psychopathology almost never require anything of a subject other than a verbal report. By contrast, ability tests almost always require subjects to do something, solve a problem, complete a task, or whatever. Wallace (1966) suggested that it might be useful to think of traits as abilities, and following that lead might very well have expanded the views of those interested in furthering clinical assessment.

4.01.7 THE ORIGINS OF CLINICAL ASSESSMENT

The earliest interest in clinical assessment was probably that used for the classification of the “insane” and mentally retarded in the early 1800s. Because there was growing interest in understanding and implementing the humane treatment of these individuals, it was first necessary to distinguish between the two types of problems. Esquirol (1838), a French physician, published a two-volume document outlining a continuum of retardation based primarily upon language (Anastasi, 1988).

Assessment in one form or another has been part of clinical psychology from its beginnings. The establishment of Wundt’s psychological laboratory at Leipzig in 1879 is considered by many to represent the birth of psychology. Wundt and the early experimental psychologists were interested in uniformity rather than assessment of the individual. In the Leipzig lab, experiments investigated psychological processes affected by perception, in which Wundt considered individual differences to be error. Accordingly, he believed that since sensitivity to stimuli differs, using a standard stimulus would compensate and thus eliminate individual differences (Wundt, Creighton, & Titchener, 1894/1896).

4.01.7.1 The Tradition of Assessment in Psychology

Sir Francis Galton’s efforts in intelligence and heritability pioneered both the formal testing movement and field testing of ideas. Through his Anthropometric Laboratory at the International Exposition in 1884, and later at the South Kensington Museum in London, Galton gathered a large database on individual differences in vision, hearing, reaction time, other sensorimotor functions, and physical characteristics. It is interesting to note that Galton’s proposi-

tion that sensory discrimination is indicative of intelligence continues to be promoted and investigated (e.g., Jensen, 1992). Galton also used questionnaire, rating scale, and free association techniques to gather data.

James McKeen Cattell, the first American student of Wundt, is credited with initiating the individual differences movement. Cattell, an important figure in American psychology, (Fourth president of the American Psychological Association and the first psychologist elected to the National Academy of Science) became interested in whether individual differences in reaction time might shed light on consciousness and, despite Wundt’s opposition, completed his dissertation on the topic. He wondered if, for example, some individuals might be observed to have fast reaction time across situations and supposed that the differences may have been lost in the averaging techniques used by Wundt and other experimental psychologists (Wiggins, 1973). Cattell later became interested in the work of Galton and extended his work by applying reaction time and other physiological processes as measures of intelligence. Cattell is credited with the first published reference to a mental test in the psychological literature (Cattell, 1890).

Cattell remained influenced by Wundt in his emphasis on psychophysical processes. Although physiological functions could be easily and accurately measured, attempts to relate them to other criteria, however, such as teacher ratings of intelligence and grades, yielded poor results (Anastasi, 1988).

Alfred Binet conducted extensive and varied research on the measurement of intelligence. His many approaches included measurements of cranial, facial, and hand form, handwriting analysis, and inkblot tests. Binet is best known for his work in the development of intelligence scales for children. The earliest form of the scale, the Binet–Simon, was developed following Binet’s appointment to a governmental commission to study the education of retarded children (Binet & Simon, 1905). The scale assessed a range of abilities with emphasis on comprehension, reasoning, and judgment. Sensorimotor and perceptual abilities were relatively less prominent, as Binet considered the broader process, for example, comprehension, to be central to intelligence. The Binet–Simon scale consisted of 30 problems arranged in order of difficulty. These problems were normed using 50 3–11-year-old normal children and a few retarded children and adults.

A second iteration, the 1908 scale, was developed. The 1908 scale was somewhat longer and normed on approximately 300 3–13-year-old normal children. Performance was grouped

by age according to the level at which 80–90% of the normal children passed, giving rise to the term “mental age.”

The Binet–Simon has been revised, translated, and adapted in numerous languages. Perhaps the most well-known revision was directed by Lewis Terman (1916) at Stanford University and this test is what is known as the Stanford–Binet. The Stanford–Binet was the origin of the intelligence quotient (IQ), the ratio between chronological and mental ages.

4.01.7.1.1 Witmer

Lightner Witmer, who studied with both Cattell and Wundt, established the first American psychological clinic at the University of Pennsylvania in 1896. This event is considered by many as the beginning of clinical psychology (Garfield, 1965; McReynolds, 1987, 1996). Witmer’s approach to assessment was focused on determining the causes of children’s problems and then to make recommendations for treatment. Diagnoses, *per se*, were not considered important, however, Witmer did make use of the Stanford–Binet and other formal assessment tools. McReynolds (1996) noted that Witmer strongly emphasized both direct observation and extensive background data as especially important for assessment.

Although Witmer characterized his work as practical, he remained committed to a scientific basis for psychology (McReynolds, 1996). It seems reasonable to conclude that Witmer was interested in assessment for bounded inference and prediction. That is, he wanted information as it might relate to specific problems for the express purpose of treating those problems (Witmer, 1996/1907).

4.01.7.1.2 Army Alpha

Robert M. Yerkes initiated and administered a program to test 1.75 million army recruits during World War I. This program, which Yerkes developed in conjunction with Terman and H. H. Goddard, administered the Army Alpha written mental test to recruits. Illiterate recruits and those failing the Alpha were given a picture-based test called the Army Beta.

Yerkes hoped that the army could be “engineered” by classifying the intelligence and capabilities of all recruits. To that end, recruits were graded from A through E and Yerkes recommended that they be assigned rank and tasks according to their tested ability. Although the army did not use the results uniformly, in many instances recruits for officer training were required to have an A or B grade on the Alpha. The tests results were later used in

controversial ways by both Yerkes and E. G. Boring to assess average American intelligence levels (see Yerkes, 1921, 1941). Despite whatever controversy may have arisen over the years, the army continues to use testing to assess aptitudes (Jensen, 1985).

4.01.8 THE RORSCHACH INKBLOT TECHNIQUE AND CLINICAL PSYCHOLOGY

The history of the Rorschach Inkblot Technique is in many ways a reflection of the history of clinical psychology in America. Clinical psychology continues to struggle with competing world views focusing on the nature of reality, the mind, and human behavior. In clinical psychology the debate about how to view the mind and behavior is usually expressed, broadly speaking, as poles of a dimension anchored by only observable behavior at one end, the influences of conscious mental processes (i.e., cognition) more in the center, and unconscious mental processes anchoring the other end. The relative importance of observable behavior and unconscious mental processes alternate with the intellectual fashions of the times.

The role of the clinical psychologist as scientist, diagnostician, and therapist continue to change, with a growing fracture between the scientifically and the clinically oriented. A central focus of debate has to do with molar vs. molecular views of personality and the ways in which personality is assessed. Conflict over the use of the Rorschach is characteristic of the debate and perturbing in light of long-standing doubts about the psychometric adequacy and the clinical usefulness of the instrument. An additional factor in the ongoing conflict in psychology seems to be that in psychology, alas, like old soldiers, theories never die. Even if refuted, they are not replaced, they only very gradually fade away (Meehl, cited by Lykken, 1991).

4.01.8.1 The Social and Philosophical Context for the Appearance of the Rorschach

Although the Rorschach was first introduced in the United States in 1925, it was during the 1940s and 1950s that the Rorschach rose to prominence in clinical psychology. The prevailing theoretical views in American academic psychology during the early years of the Rorschach were Gestalt and behaviorism. In many ways the interest and devotion of Rorschach proponents to the technique seems to have been a reaction against what they saw as reductionist and positivistic approaches to

personality assessment on the part of behaviorists and often atheoretical psychometricians. Additionally, behaviorists focused on environmental determinants of behavior at the same time that psychoanalytic theory, in spite of its rejection in much of academia, was beginning to flourish in clinical psychology. Moreover, by the late 1940s, many psychologists were interested in reviving the notion of the self, which had been rejected by behaviorism and psychoanalysis (Reisman, 1991).

Proponents of the Rorschach believed that underlying dimensions of “true” personality could be elicited only by indirect, projective methods; defense mechanisms, repression, and perhaps other unconscious processes prevented an individual from having access to critical information about him- or herself. Direct assessment of personality was narrow and incomplete, but the ambiguity of the inkblot stimulus material would elicit true responses. Because during the 1940s and 1950s testing was virtually the only applied professional activity performed by clinical psychologists (Millon, 1984), it is not surprising that the Rorschach would generate a great deal of interest and activity. What is surprising is that a test criticized even then and continuously until now as being too subjective in administration, scoring, and interpretation, of questionable reliability, and of dubious validity, would be continually used for 70 years.

Rorschach proponents did claim to view the technique as scientific, and there were attempts to establish norms and to approach the Rorschach scientifically, but we view the Rorschach ultimately as what Richard Feynman (1986) refers to as “Cargo Cult Science:”

In the South Seas there is a cargo cult of people. During the war, they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to make things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit on, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks just the way it looked before. But it doesn't work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent percepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.

The Rorschach technique is missing something essential. Although as we stated earlier, people almost certainly “project” aspects of their personality on to ambiguous stimuli, use of the Rorschach has failed to demonstrate con-

vincing evidence of validity in decades of attempts to find it. The planes still don't land.

4.01.8.2 The Birth of the Rorschach

Whether and how to use the Rorschach has been a source of controversy since its introduction. Perhaps much of the controversy and dissent about scoring and interpretation of responses to the inkblots among advocates of the technique were a result of its founder's death a few months after the publication of his initial monograph detailing 10 years of studies with inkblots, leaving a nascent method open to various interpretations. The original notions of using the technique tentatively and experimentally began fading with its founder's death, being replaced by an overriding concern for clinical uses.

Herman Rorschach, the son of a Swiss art teacher, began experimenting with inkblots in various psychopathic hospitals in 1911, the year after completing his medical training (Klopfer & Kelley, 1942). The Rorschach method was introduced in the United States in 1925 by David Levy, a psychologist and psychiatrist (Hertz, 1986; Klopfer & Kelley, 1942), who had been a student of Emil Oberholzer, Rorschach's closest medical colleague and who continued Rorschach's work after his death. Levy taught the technique to Samuel Beck, who wrote his dissertation on the technique and published the first manual on the Rorschach in 1937 (Exner, 1969; Hertz, 1986; Klopfer & Kelley, 1942). Beck and Bruno Klopfer were probably the most influential individuals in terms of widening the use of the technique, as well as in fomenting debate about how to score and interpret Rorschach responses. Beck was more behavioral and experimental in his approach and strongly advocated establishing norms and testing the validity of responses. Klopfer, a German who had studied with Jung in Switzerland after fleeing Hitler and before coming to the United States, was much more inferential in his interpretation and scoring. Rorschach himself was considerably more tentative about his findings than subsequent proponents of the technique were, or than they seem to be to this day.

It is likely that dissemination of the Rorschach was actually helped by the controversy and dissent within the ranks of Rorschach adherents, as well as by the fight against perceived rigid standards of psychometrics and nomothetic personality theories. The internal debate among adherents of various systems of scoring and interpretation seemed to foster beliefs that the findings finally proving them right were just around the corner. This

belief of imminent justification seems to characterize even present day Rorschach proponents.

Another faction of Rorschach adherents with more interest in applying the Rorschach to clinical cases took the view that assessment and prediction based on clinical judgment and acumen are inherently superior to psychometric and statistical assessment and prediction. During the 1950s and 1960s, the emphasis shifted from scores and scoring systems to the utilization of clinical acumen and sensitivity, and attempts to understand subtle aspects of the entire testing situation (Sarason, 1954). As the role of the clinical psychologist expanded into more applied clinical activity, practitioners' attention to the experimental scientific roots of the discipline began fading. With this movement further from a scientific basis for theories and techniques, the theories promoted by academic psychologists were considered mechanistic by most practitioners. As a result, the academics' criticisms about projectives such as the Rorschach were increasingly viewed as invalid (or, perhaps worse, as irrelevant). In our reading of the literature, it appears that even those Rorschach supporters who believe science is important cling to the "Cargo Cult Science" of ratios and scoring systems lacking in empirical support but with the expectation of redemption almost momentarily expected.

This shift in the 1950s and 1960s to a focus on clinical skills was in the context of the emergence of psychotherapy as a primary professional activity for psychologists. Erikson's theory of psychosocial development was embraced, psychodynamic theory in various forms (Adler, Rank, Horney, Sullivan) was popular with clinicians, and Rogerian humanistic psychology emerged along with behavior modification and systematic desensitization (Reisman, 1991). In psychiatry there were rapid advances in psychotropic medications. These changes in the field seemed to steel the resolve of clinicians who believed that human psychology could not be reduced to biology, classification, and statistical formulas. Despite the lack of any demonstrated validity of the Rorschach from research studies, clinicians focused on the feedback they received, or thought they received, from clients, and believed the Rorschach helped them to better understand their clients. At about the same time as these developments, Paul Meehl (1954) published an analysis of the general problem of clinical vs. statistical prediction.

4.01.8.3 Clinical vs. Statistical Prediction

The central issue in relation to comparisons of clinical and statistical (actuarial) prediction is

simple: when there is a database of cases with known outcome, can a skilled clinician use his or her judgment to combine the relevant information about a client into the correct formulations (predictions) as well as or better than a statistical formula that uses the same information? The answer, based on numerous studies in which clinicians had as much or more information as was entered into the statistical prediction, is *no*. Clinicians occasionally equal but never exceed statistical predictions of behavior, diagnoses, psychotherapy outcome, and like events of interest. The preponderance of evidence favors statistical prediction.

Even when statistical models are based upon the information used by clinicians, the models outperform the clinicians on whom they are based (Dawes et al., 1989). Exceptions do occur in circumstances of events that reverse the actuarial formula or of judgments mediated by theories that are, therefore, difficult or even impossible to duplicate statistically (Dawes et al., 1989). When such information is available to clinicians, and those circumstances may be infrequent, they are likely to outperform statistical models. Meehl (1954) referred to these rare events as the broken leg phenomenon. That name was derived from an illustration in which a statistical formula is highly successful in predicting an individual's weekly attendance at a movie, but should be discarded upon discovering that the subject is in a cast with a fractured femur.

One reason for the superiority of statistical prediction is that clinicians tend to think that too many cases are exceptions to ordinary rules and, even in the case of rare events, they ultimately perform better when they rely strictly on statistical conclusions (Goldberg, 1968). The human mind is a poor computer and does not do a good job at quantifying and weighting observations, the very things that regression equations were invented for (Goldberg, 1991). We do not mean to suggest that statistical formulas can be used to perform psychotherapy, or that the predictions could be made without first gathering the relevant observations from clinicians. We also bear in mind that a great many clinical decisions are made in circumstances in which there is no known outcome.

The debate about clinical vs. statistical prediction has been characterized by *ad hominem* attacks, and Meehl (1954) started his book, *Clinical versus statistical prediction*, with lists of invective from both sides of the argument. Briefly, opponents of statistical prediction have suggested that the approach is atomistic, inhuman, arbitrary, and oversimplified, while its proponents suggest that it is objective, reliable,

rigorous, and scientific. Conversely, negative appraisals of clinical prediction suggest the method is sloppy, muddleheaded, unscientific, and vague, while its proponents suggest that the method is dynamic, sensitive, meaningful, and holistic (Meehl, 1954).

The case for the use of psychodiagnostic tests such as the Rorschach and the validity of clinical observation of relationships between thoughts, behavior, and personality characteristics becomes questionable considering the findings about the questionable validity of clinical judgments. Further, it has been known for a long while that statements from clinical assessments and psychological reports are often of universal applicability (Forer, 1949). When previously prepared statements representative of those in psychological evaluations are presented to a variety of individuals, the individuals enthusiastically agree that the statements uniquely apply to them. Therefore, it seems that the very evidence often used by clinicians, that their clients believe assessments to be accurate and that they are helped by assessment and treatment, affords no reassurance. Much information provided by typical psychodiagnostic feedback is general and applies to almost anyone. The associations with various personality characteristics, signs, and indicators may be more related to what any astute observer has learned to associate with them through observation, folklore, and literature, that is, "illusory correlations" (Chapman & Chapman, 1967, 1969; Reisman, 1991). It is likely that such illusory correlations are involved in accounts of individuals known as "Rorschach Savants," who are purported anecdotally to see phenomenal amounts of information in Rorschach responses.

It is astonishing that the Rorschach continues to be not only very popular, but in many states is required as part of forensic psychological assessment in child custody disputes (Robyn Dawes, personal communication). Reisman (1991) suggests that the failure of clinical psychologists to modify their behavior no matter how much aversive stimulation is applied is less a refutation of Skinner's theory than evidence of a great capacity to distort information. Many clinicians and even some researchers continue to believe in the validity of the Rorschach (and other projective tests) in spite of overwhelming evidence to the contrary and almost universal agreement among the scientific community that the central assumption on which the Rorschach is based is faulty.

The entire Rorschach is based on a fallacious assumption, namely that indirect (projective) methods are more valid than direct (self-rating

or questionnaire) methods because people are so repressed that they cannot describe their real emotions and impulses. A large body of literature indicates the fallacy of this assumption. Even within self-report items, more content obvious items prove to be more valid than subtle ones. Why give an hour test, with another hour to score, to get a crude estimate of anxiety or depression which is usually less reliable and valid than a short true-false scale which takes a few minutes and where there is no unreliability of scoring? I have compared direct and indirect (Rorschach and TAT) measures of dependency, anxiety, depression, and hostility using peer ratings as criteria. The most indirect methods have zero validity, the most direct methods have low to moderate validity, and methods which are intermediate in directness (e.g., sentence completion) are intermediate in validity. A great deal of effort was expended in scoring content from TAT and Rorschach and consensus agreement was obtained where disagreement in scoring occurred. All this was to no avail because the two projectives did not correlate with each other let alone with the criteria or any of the direct methods. (Marvin Zuckerman, SSCPNET, April 22, 1996)

Although yet another scoring system for the Rorschach has been used and researched for the past 20 years (Exner, 1974, 1993) with a greater emphasis on standardization of scoring and interpretation, it has yielded no apparent improvement on the predictive or incremental validity of the technique. Criticisms of the research are nearly identical to those expressed in the 1940s and 1950s. Disturbingly, in spite of overwhelming evidence of their invalidity, clinicians tend to continue to rely on their impressions and interpretations of the content of Rorschach responses (Reisman, 1991). It is not precisely fair to say that the Rorschach is unrelated to anything, but its validity is so limited as to leave virtually no real utility for its use. Most problematic, it is inferior to and more time-consuming than instruments with better reliability and validity and the Rorschach appears to have zero incremental validity (Sechrest, 1963).

4.01.8.4 Old Tests Never Die, They Just Fade Away

The continued drift of psychology away from its scientific roots does not appear to be slowing. This drift seems additionally fueled by economic and employment concerns and continued training of too many practitioners. The current conflict is unlikely to slow as managed health care and cutbacks in federal funding lessen job opportunities, and the future of psychology is uncertain. Clinical psychology, even in the halcyon days of the scientist-practitioner model,

was never resolute in its commitment to science. For example, students coming into the field were generally not required to have any particular prior training in science, or its principal handmaiden, mathematics, and they needed only to declare a personal fealty to the idea of research. That situation has almost certainly become much worse over the past two or three decades of drift toward practitioner–scientist, then practitioner–scholar, and then frankly practitioner programs. The net result is that clinical psychology has a huge number of practitioners who are not only ill-equipped to handle the demands of evaluating the scientific basis for practice, but they are ill-disposed even to doing so. Economic pressures and their own incapacities make scientific evidence, which is at best likely to be disappointing, a threat. Anecdotes, “clinical experience,” and so on are far more reassuring and, hence, attractive. Better to believe in an unproven instrument or procedure than to be deprived of any basis for pride and survival.

Lykken (1991) noted that present knowledge in psychology is very broad but very shallow. Most recently trained clinical psychologists probably have little acquaintance with the philosophy of science and not much knowledge of the clinical vs. statistical prediction literature; certainly they have inadequate training in measurement, statistics, and probability. This ignorance of the roots of psychological theory and scientific psychology contributes to the continued use of a completely unjustifiable procedure such as the Rorschach. It is difficult to refute disproven techniques and theories when a class of the profession basis its identity and livelihood on them. The problem of theories fading away and reviving as suggested by Meehl’s “old soldiers” simile is not restricted to clinical psychology; psychology as a whole operates in this way.

In other sciences, each generation builds on the foundations of the discipline’s previous scientists. Psychology seems to view its predecessors as “intrepid explorers who came back empty-handed” (Lykken, 1991). To be fair, establishing a psychological science is extremely difficult because it is difficult to operationalize psychological constructs and because there is notable measurement error. The profession and practice of clinical psychology would be helped immensely, however, if we could better educate graduate students in philosophy of science, measurement, and statistics, in addition to psychological theory.

The Rorschach did not come into prominence originally because of evidence for its superiority over existing measures, for example, questionnaires and checklists. It was adopted eagerly, we

think, more because of discontent with the obvious inadequacies of existing alternatives. We suspect that whatever its own inadequacies, the Rorschach will not die but will only fade away when some alternative instrument or procedure becomes available and seems potentially to be a better one.

4.01.9 OTHER MEASURES USED IN CLINICAL PSYCHOLOGY

The list of measures that have been used in clinical psychology is very long, and many appear simply to have faded away. For example, two projective tests that once had a spate of popularity are the Blacky Test and the Make-a-Picture Story Test (MAPS) (Shneidman, 1986). The Blacky Test seems to have disappeared altogether, and the MAPS is rarely encountered in the literature. Neither was ever demonstrated to be less reliable or less valid than other tests; each simply appears to have faded away, the Blacky probably because its version of psychoanalytic theory has also faded somewhat and the MAPS because it was cumbersome and slow to administer. There is not much point in recounting the histories of the many now deservedly (even if not uniquely deserved) forgotten tests.

4.01.9.1 The Thematic Apperception Test

Morgan and Murray (1935) introduced the Thematic Apperception Test (TAT) based on what they termed the “well-recognized fact” that when presented with ambiguous stimuli people reveal their own personality. The TAT consists of a series of pictures of ambiguous social situations in which the examinee describes the social situation as they see it. The TAT was originally designed to be interpreted in light of psychoanalytic theory, the theory driving its design. There were subsequently a variety of scoring systems from different perspectives, although none has improved on the recurrent problem of inconsistency in use from clinician to clinician.

The TAT, as one might imagine, can be scored more or less reliably, depending on the nature of the variable involved and the adequacy of its definition. The major problem is what the scores may be related to and how they may be interpreted. Over the many years of its existence, TAT scores have been related to many different phenomena, sometimes with moderate success. The literature would show that achievement has been extensively studied by way of the TAT (see Keiser & Prather, 1990) as have other needs or motives. Although the

research is reasonably consistent in showing some evidence for validity of some TAT scores and the instrument has proven to be of some value in research, the evidence was never strong enough to justify use of the TAT for individual decision-making in clinical settings. The TAT, like most other clinical measures, can at best be considered enlightening.

4.01.9.2 Sentence Completion Tests

Another variety of quasiprojective instruments is the sentence completion test, which consists of a stem, for example, "When I was a child," that the respondent is supposed to make into a complete sentence by writing down his or her own thoughts. The sentence completion test, of which the Rotter Incomplete Sentences Blank (Rotter & Rafferty, 1950) is the best known version, probably evolved from word association tests, which go back to Galton, Cattell, and Kraepelin in the latter part of the nineteenth century (Anastasi, 1988). The Rotter ISB was considered to be a measure of psychological conflict and, therefore, adjustment, and like so many other measures, under the right circumstances, it could be scored in a reasonably dependable way and could result in "significant" validity coefficients. That is to say, the ISB could be shown variously and not invariably to be correlated around 0.30 with criteria thought by someone to be of interest. Those correlations might be useful for some research purposes, but they were not grounds for much confidence in clinical settings. They may, however, in the minds of many clinicians have inspired more confidence and, therefore, more use than was warranted.

4.01.9.3 Objective Testing

The term "objective test" usually refers to a self-report measure that presents a stimulus item to a respondent and that requires a constrained response such as "True/False," "Agree/Disagree," and so forth. There are many, many objective tests, but the dominant one is, and virtually always has been, the MMPI (Hathaway & McKinley, 1943). We have already discussed various aspects of the MMPI under other topics, but it is worth noting here that the durability of the MMPI has been impressive. Its clinical utility has not. It yields profiles that seem impressive, and it certainly can, in general, serve as a screening instrument for psychopathology: people who get really high scores on one or more of the MMPI scales probably have *something* awry in their lives. No relationships have ever been consistently demonstrated

between the MMPI and functional capacities or incapacities that would justify clinical decisions other than to seek further information about the client or patient.

The MMPI more than other available instruments has been automated, to the extent of producing computer-based interpretations of test profiles. An unfortunate limitation of computer-based interpretations is that, because of their proprietary nature, the algorithms underlying them are not available. Consequently, one cannot know which interpretations are based on empirical evidence and which, perhaps, on clinical lore, let alone how good the evidence might be. Such interpretations must be accepted on faith. When the MMPI is used in a fully automatic mode, it is questionable whether it even should be considered a clinical assessment.

4.01.9.4 The Clinician as a Clinical Instrument

Clinical psychology has never been completely clear about whether it wishes to distinguish between the test—a tool—and the test-in-the-hands-of-a-user. The perspective of standardized testing implies that the test is a tool that, in the hands of any properly trained user, should produce the same results for any given examinee. Many clinical instruments, however, cannot be considered to be so tightly standardized, and it is to be expected that results might differ, perhaps even substantially, from one examiner to another, even for the same examinee. Within reason, at least, an examinee's performance on a vocabulary test or a trail-making test should be little affected by the characteristics of the examiner, nor should the scoring and interpretation of the performance. By contrast, an examinee's responses might be affected to a considerable degree by the characteristics of an examiner administering a Rorschach or a TAT, let alone the interpretation of those responses.

The field of clinical psychology abounds in tales of diagnostic acumen of marvelous proportions manifested by legendary clinicians able to use the Rorschach, an MMPI profile, or some other instrument as a stimulus. Unfortunately, no such tales have advanced beyond the bounds of anecdote, and none of these legendary clinicians appears to have been able to pass along his or her acumen to a group of students—let alone passing it along across several generations. Consequently, if clinicians are to be part of the clinical assessment equation, then it seems inevitable that individual clinicians will have to be validated individually, that is, individual clinicians will

have to be shown to be reliable and valid instruments. That will not further progress in the field.

4.01.9.5 Structured Interviews

A fairly recent development in clinical assessment is the structured interview schedule. These schedules are intended to produce a diagnostic judgment related to the *DSM* (American Psychiatric Association, 1994), a narrow, bounded purpose. There are several such interview schedules currently available, but we will discuss the Structured Clinical Interview for *DSM-IV* (SCID) as an example and because it is probably the one most widely used.

As noted earlier, most psychological assessment appears to be done for purposes of enlightenment rather than for decision-making. Nevertheless, diagnoses are often required for reimbursement, medication referrals, custody evaluations, and forensic assessments. The SCID (Spitzer, Gibbon, & Williams, 1997) appears to be used quite infrequently in other than research settings, for example, it is not mentioned on any list of instruments used by clinicians. That neglect is interesting in view of the attention that was paid to the development of the SCID and its established dependability.

Use of the SCID in clinical practice would probably contribute to improved assessment (and presumably to more appropriate treatment), whether for specific *DSM* diagnostic purposes or simply for gathering pertinent information. The SCID was designed to capitalize on clinical skills and to be more "clinician-friendly" than other structured interviews (Spitzer, Williams, Gibbon, & First, 1992). The SCID is meant to be used by precisely those people who can already conduct an interview, and although the SCID is somewhat time-consuming, but probably less so than, say, the Rorschach, psychologists interview all patients, and for most clinicians to do so in a structured manner would not be a significant departure. That is, the time would be spent interviewing the patient and the SCID would not add much if anything in terms of time or cost to standard practice. The SCID demonstrates good reliability (test-retest and inter-rater) for most disorders, with kappa coefficients averaging 0.60–0.80 or greater (Segal, Hersen, & Van Hasselt, 1995; Williams, Gibbon, First, & Spitzer, 1992). Agreement between diagnoses obtained by SCID and by traditional clinical interviews is poor to moderate with average kappa coefficients of 0.25 (Steiner, Tebes, Sledge, & Walker, 1995), suggesting strongly that reliance on unstructured clinical interviews is unwise.

Perhaps the SCID is not used because it takes some training and practice to become proficient in its use. That requirement is certainly different from the typical assessment instruments advertised in psychological publications, which boast their quick and easy use and say nothing about their reliability and validity. It may also be that beliefs about the superiority of clinical judgment over other more structured practices, for example, the use of projective tests, contributes strongly as well. Whatever the reasons for lack of clinical use of the SCID, and we suspect that it is both training time and beliefs about clinical skill, it is an unfortunate omission from assessment practice.

4.01.10 CONCLUSIONS

Progress in psychological assessment, at least for clinical applications, has been disappointing over the century since the field started. Conceptual and theoretical developments have been minimal, although we might except some observational methods used primarily in behavioral work and some research settings. The field continues to move away from its scientific roots in psychology, and clinical assessment has no other base on which to build any conceptual structure. Moreover, clinical assessment has never been more than minimally guided by psychometric theory and analysis, for example, scarcely beyond superficial concern with "reliability" of measures, and graduate education and training in research methods and measurement is at an ebb and is maybe still decreasing. Overall, clinical assessment as an enterprise seems to be cut adrift from any important sources of rigor, and almost anything goes. Perhaps it is fortunate, then, that despite the frequent insistence on assessment as a cornerstone of the practice of clinical psychology, there is much less evidence for its importance and prevalence than would be expected.

4.01.11 REFERENCES

- Achenbach, T.M. (1985). *Assessment and taxonomy of child and adolescent psychopathology*. Beverly Hills, CA: Sage.
- Achenbach, T.M., & Edelbrock, C. S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: Department of Psychiatry, University of Vermont.
- Achenbach, T. M., & Edelbrock, C. S. (1986). *Manual for the Teachers Report Form and Teacher Version of the Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., & Edelbrock, C. S. (1987). *Manual for the Youth Self-Report Form and Youth Version of the Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Aiken, L., West, S. G., Sechrest, L., & Reno, R. (1990). Graduate training in statistics, methodology, and

- measurement in psychology: A survey of Ph.D. programs in North America. *American Psychologist*, 45, 721–734.
- American Psychiatric Association (1994). *Diagnostic and statistical manual for mental disorders* (4th ed.). Washington, DC: Author.
- American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, DC: Author.
- APA Practice Directorate (1996). Practitioner survey results offer comprehensive view of psychological practice. *Practitioner Update*, 4(2).
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Atkins, M. S., Pelham, W. E., & White, K. J. (1990). Hyperactivity and attention deficit disorders. In M. Hersen & V. B. Van Hasselt (Eds.), *Psychological aspects of developmental and physical disabilities: A casebook*. Newbury Park, CA: Sage.
- Berg, I. A. (1961). *Measuring deviant behavior by means of deviant response sets*. New York: Harpers.
- Bergen, A. E., & Garfield, S. L. (1994). *Handbook of psychotherapy and behavior change*. New York: Wiley.
- Bernreuter, R. G. (1933). Validity of the personality inventory. *Personality Journal*, 11, 383–386.
- Binet, A., & Simon, T. H. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Année Psychologique*, 11, 191–244.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cattell, J. M. (1890). Mental tests and measurements. *Mind*, 15, 373–380.
- Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, 193–204.
- Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271–280.
- Costello, A., Edelbrock, C. S., Kalas, R., Kessler, M., & Klaric, S. A. (1982). *Diagnostic Interview Schedule for Children (DISC)*. Bethesda, MD: National Institute for Mental Health.
- Craik, K. H. (1986). Personality research methods: An historical perspective. *Journal of Personality*, 54(1), 18–51.
- Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper and Row.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137–163.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–1674.
- Epstein, S. (1983). Aggregation and beyond: Some basic issues in the prediction of behavior. *Journal of Personality*, 51, 360–392.
- Esquirol, J. E. D. (1838). *Des maladies mentales considerees sous les rapports medical, hygienique, et medico-legal* (2 Vols.). Paris: Bailliere.
- Exner, J. E. (1969). *The Rorschach systems*. New York: Grune & Stratton.
- Exner, J. E. (1974). *The Rorschach systems*. New York: Grune & Stratton.
- Exner, J. E. (1986). *The Rorschach: A comprehensive system*. New York: Wiley.
- Exner, J. E. (1993). *The Rorschach: A comprehensive system*. New York: Wiley.
- Feynman, R. (1986). *Surely you're joking, Mr. Feynman!* New York: Bantam Books.
- Fiske, D. W. (1971). *Measuring the concepts of personality*. Chicago: Aldine Press.
- Forer, B. (1949). The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology*, 44, 118–123.
- Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology*, 8, 389–413.
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–338). Hillsdale, NJ: Erlbaum.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23, 483–496.
- Goldberg, L. R. (1991). Human mind versus regression equation: Five contrasts. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. 1. Matters of public interest: essays in honor of Paul E. Meehl* (pp. 173–184). Minneapolis, MN: University of Minnesota Press.
- Goyette, C. H., Conners, C. K., & Ulrich, R. E. (1978). Normative data on the Conner's parent and teacher rating scales. *Journal of Abnormal Child Psychology*, 6(2), 221–236.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, & Law*, 2(2), 293–323.
- Hathaway, S. R., & McKinley, M. N. (1943). *The Minnesota Multiphasic Personality Inventory* (Rev. ed.). Minneapolis, MN: University of Minnesota Press.
- Hertz, M. R. (1986). Rorschachbound: A 50-year memoir. *Journal of Personality Assessment*, 50(3), 396–416.
- Hoza, B., Vallano, G., & Pelham, W. E. (1995). Attention-deficit/hyperactivity disorder. In R. T. Ammerman & M. Hersen (Eds.), *Handbook of child behavior therapy in psychiatric setting*. New York: Wiley.
- Hunt, W. C. (1956). *The clinical psychologist*. Springfield, IL: Thomas.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Jackson, D. N., & Messick, S. (1962). Response styles on the MMPI: Comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology*, 65, 285–299.
- Jensen, A. R. (1985). Description & utility of Armed Services Vocational Aptitude Battery-14. *Measurement & Evaluation in Counseling & Development*, 18(1), 32–37.
- Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality & Individual Differences*, 13(8), 869–881.
- Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement (K-TEA)*. Circle Pines, MN: American Guidance Service.
- Kagan, J. (1959). The stability of TAT fantasy and stimulus ambiguity. *Journal of Consulting Psychology*, 23, 266–271.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
- Keiser, R. E., & Prather, E. N. (1990). What is the TAT? A review of ten years of research. *Journal of Personality Assessment*, 55 (3–4), 800–803.
- Klopper, B., & Kelley, D. M. (1942). *The Rorschach technique*. Yonkers-on-Hudson, NY: World Book Company.
- Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage.

- Levy, L. H. (1963). *Psychological interpretation*. New York: Holt, Rinehart, and Winston.
- Lindzey, G. (1952). Thematic Apperception Test: Interpretive assumptions and related empirical evidence. *Psychological Bulletin*.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, No. 7.
- Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology* (pp. 3–39). Minneapolis, MN: University of Minnesota Press.
- Maloney, M. P., & Ward, M. P. (1976). *Psychological assessment: A conceptual approach*. New York: Oxford University Press.
- McClure, D. G., & Gordon, M. (1984). Performance of disturbed hyperactive and nonhyperactive children on an objective measure of hyperactivity. *Journal of Abnormal Child Psychology*, 12(4), 561–571.
- McCracken, B. A., & McCallum, S. R. (1993). *Wechsler Intelligence Scale for Children* (3rd ed.). Brandon, VT: Clinical Psychology Publishing.
- McReynolds, P. (1987). Lightner Witmer: Little known founder of clinical psychology. *American Psychologist*, 42, 849–858.
- McReynolds, P. (1996). Lightner Witmer: A centennial tribute. *American Psychologist*, 51(3), 237–240.
- Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis, MN: University of Minnesota Press.
- Meehl, P. E. (1960). The cognitive activity of the clinician. *The American Psychologist*, 15, 19–27.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.
- Meier, S. L. (1994). *The chronic crisis in psychological measurement and assessment: A historical survey*. New York: Academic Press.
- Meyer, G. J. (1996). The Rorschach and MMPI: Toward a more scientific differential understanding of cross-method assessment. *Journal of Personality Assessment*, 67, 558–578.
- Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment*, 69, 1–38.
- Millon, T. (1984). On the renaissance of personality assessment and personality theory. *Journal of Personality Assessment*, 48(5), 450–466.
- Millon, T., & Davis, R. D. (1993). The Millon Adolescent Personality Inventory and the Millon Adolescent Clinical Inventory. *Journal of Counseling and Development*.
- Mitchell, J. V., Jr. (Ed.) (1985). *The mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska.
- Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies. *Archives of Neurological Psychiatry*, 35, 289–306.
- Murray, H. A. (1943). *Manual for the Thematic Apperception Test*. Cambridge, MA: Harvard University Press.
- Murstein, B. I. (1963). *Theory and research in projective techniques*. New York: Wiley.
- Prochaska, J. O., DiClemente, C. C., & Norcross, J. C. (1992). In search of how people change: Applications to addictive behaviors. *American Psychologist*, 47(9), 1102–1114.
- Rotter, J. B., & Rafferty, J. E. (1950). *Manual: The Rotter Incomplete Sentences Blank*. San Antonio, TX: Psychological Corporation.
- Reisman, J. M. (1991). *A history of clinical psychology* (2nd ed.). New York: Hemisphere.
- Samejima, F. (1988). Comprehensive latent trait theory. *Behaviormetrika*, 24, 1–24.
- Samelson, F. (1987). Was early mental testing (a) racist inspired, (b) objective science, (c) a technology for democracy, (d) the origin of multiple-choice exams, (e) none of the above? (Mark the RIGHT answer). In M. M. Sokal (Ed.) *Psychological testing and American society 1890–1930* (pp. 113–127). New Brunswick, NJ: Rutgers University Press.
- Sarason, S. B. (1954). *The clinical interaction, with special reference to the Rorschach*. New York: Harper.
- Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, 33(1), 153–158.
- Sechrest, L. (1968). Testing, measuring, and assessing people. In W. W. Lambert & E. G. Borgatta (Eds.), *Handbook of personality theory and research*. Chicago: Rand McNally.
- Sechrest, L. (1992). The past future of clinical psychology: A reflection on Woodworth (1937). *Journal of Consulting and Clinical Psychology*, 60(1), 18–23.
- Sechrest, L., McKnight, P. E., & McKnight, K. M. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist*, 51, 1065–1071.
- Segal, D. L., Hersen, M., & Van Hasselt, V. B. (1994). Reliability of the structured clinical interview for DSM-III-R: An evaluative review. *Comprehensive Psychiatry*, 35(4), 316–327.
- Sharkey, K. J., & Ritzler, B. A. (1985). Comparing the diagnostic validity of the TAT and a New Picture Projective Test. *Journal of Personality Assessment*, 49, 406–412.
- Shneidman, E. S. (1986). MAPS of the Harvard Yard. *Journal of Personality Assessment*, 50(3), 436–447.
- Somoza, E., Steer, R. A., Beck, A. T., & Clark, D. A. (1994). Differentiating major depression and panic disorders by self-report and clinical rating scales: ROC analysis and information theory. *Behaviour Research and Therapy*, 32, 771–782.
- Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV Disorders (SCID-I)-Clinician Version*. Washington, DC: American Psychiatric Press.
- Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1992). The Structured Clinical Interview for DSM-III-R (SCID): I. History, rationale, and description. *Archives of General Psychiatry*, 49(8), 624–629.
- Steiner, J. L., Tebes, J. K., Sledge, W. H., & Walker, M. L. (1995). A comparison of the Structured Clinical Interview for DSM-III-R and clinical diagnoses. *Journal of Nervous & Mental Disease*, 183(6), 365–369.
- Strupp, H. H., Horowitz, L. M., & Lambert, M. J. (1997). *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery*. Washington, DC: American Psychological Association.
- Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.
- Thorndike, R., & Hagen, E. (1955). *Measurement and evaluation in psychology and education*. New York: Wiley.
- Wade, T. C., & Baker, T. B. (1977). Opinions and use of psychological tests: A survey of clinical psychologists. *American Psychologist*, 32, 874–882.
- Wallace, J. (1966). An abilities conception of personality: Some implications for personality measurement. *American Psychologist*, 21(2), 132–138.
- Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): 1. Conceptual framework and item selection. *Medical Care*, 30(6), 473–483.
- Watkins, C. E. (1991). What have surveys taught us about the teaching and practice of psychological assessment? *Journal of Personality Assessment*, 56, 426–437.
- Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice*, 26, 54–60.

- Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally.
- Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., & Grove, J. B. (1981). *Nonreactive measures in the social sciences*. Boston: Houghton Mifflin.
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.
- Williams, J. B. W., Gibbon, M., First, M. B., & Spitzer, R. L. (1992). The Structured Clinical Interview for DSM-III-R (SCID): II. Multisite test-retest reliability. *Archives of General Psychiatry*, 49(8), 630-636.
- Witmer, L. (1996). Clinical Psychology. *American Psychologist*, 51(3), 248-251. (Original work published 1907.)
- Woodworth, R. S. (1992). The future of clinical psychology. *Journal of Consulting and Clinical Psychology*, 60, 16-17. (Original work published 1937.)
- Wundt, W., Creighton, J. E., & Titchener, E. B. (1894/1896). *Lectures on human and animal psychology*. London: Swan Sonnenschein.
- Yerkes, R. M. (Ed.) (1921). Psychological examining in the United States army. *Memoirs of the National Academy of Sciences*, 15.
- Yerkes, R. M. (1941). Man power and military effectiveness: The case for human engineering. *Journal of Consulting Psychology*, 5, 205-209.
- Zuckerman, M. (1996, April 22). *Society for a Science of Clinical Psychology Network* (SSCPNET; electronic network).

4.02

Fundamentals of Measurement and Assessment in Psychology

CECIL R. REYNOLDS

Texas A&M University, College Station, TX, USA

4.02.1 INTRODUCTION	33
4.02.2 NORMS AND SCALES OF MEASUREMENT	34
4.02.2.1 <i>Scales of Measurement</i>	34
4.02.2.1.1 <i>Nominal scales</i>	34
4.02.2.1.2 <i>Ordinal scales</i>	34
4.02.2.1.3 <i>Interval scales</i>	35
4.02.2.1.4 <i>Ratio scales</i>	35
4.02.2.2 <i>Norms and Reference Groups</i>	35
4.02.3 UNITS OF MEASUREMENT	37
4.02.4 ACCURACY OF TEST SCORES	41
4.02.4.1 <i>True Score Theory</i>	41
4.02.4.2 <i>Generalizability Theory</i>	43
4.02.5 VALIDITY	43
4.02.6 THE ASSESSMENT PROCESS	45
4.02.7 MODELS AND METHODS OF ASSESSMENT	46
4.02.7.1 <i>Traditional Norm-referenced Assessment</i>	46
4.02.7.1.1 <i>Intelligence, achievement, and special abilities</i>	46
4.02.7.2 <i>Norm-referenced, Objective Personality Measures</i>	48
4.02.7.3 <i>Projective Assessment</i>	49
4.02.7.4 <i>Behavioral Assessment</i>	50
4.02.7.5 <i>Neuropsychological Assessment</i>	51
4.02.8 CLINICAL VS. STATISTICAL PREDICTION	52
4.02.9 ACCESSING CRITICAL COMMENTARY ON STANDARDIZED PSYCHOLOGICAL TESTS	53
4.02.10 CONCLUDING REMARKS	53
4.02.11 REFERENCES	54

4.02.1 INTRODUCTION

Measurement is a set of rules for assigning numbers to objects or entities. A psychological measuring device (typically a test), then, is a set of rules (the test questions, directions for administration, scoring criteria, etc.) for assigning numbers to an individual that are believed to

represent a level of some particular psychological trait, attribute, or behavior of the individual. These characteristics may be observable directly or may be inferred or observed indirectly through changes in behavior or responses to a set or a variable stimulus. Assessment is a more comprehensive process of deriving meaning from test scores and clinical

information in order to describe the individual both broadly and in depth. Psychological tests are the nonexclusive tools of assessment. A proper assessment must also consider the background and current cultural milieu of the individual and actual observed behavior. This chapter does not attempt to deal with all aspects of the assessment process. An introduction to basic measurement technology and theory will be provided along with material concerning different methods of measurement intended to enhance understanding of other chapters in this work.

There are many problems and controversial issues in psychological and educational assessment and, obviously, all cannot be treated in this work. As one example, assessment and the testing that accompanies it occur within a particular situation or context. The results that are obtained may thus be strongly influenced by situational factors in the case of some individuals but less so or not at all for others. The question of the generalizability of test results obtained under a specified set of conditions takes on major importance in interpreting test scores. Not all variables that influence generalizability are known and few that have been well researched. Test anxiety is one factor thought to influence strongly the generalizability of results across settings and has been researched extensively, yet the complete articulation of the relationship among test anxiety, test performance, and the validity of test-score interpretations across settings is far from complete. The assessment of children, in particular, poses special problems because of the rapid growth and development as well as their susceptibility to external environmental factors. Many of these factors are treated at length in Anastasi (1981), Cronbach (1983), Kaufman (1994), Reynolds (1985), and Reynolds and Kamphaus (1990a, 1990b), and the interested reader is referred to these sources for further reading on the problems, issues, and limitations of educational and psychological testing, as well as to the other chapters in this volume and to Volume 10.

4.02.2 NORMS AND SCALES OF MEASUREMENT

Many pieces of information are necessary before one can attach the proper meaning to a test score. Among the basic are knowledge of what scale of measurement has been employed and with what sort of reference group the individual is being compared, if any. Different scales have different properties and convey different levels and types of information just as

they do in other arenas; for example, four inches of water conveys a very different meaning than a reference to four gallons of water. The four basic scales of measurement are nominal, ordinal, interval, and ratio scales. As one moves from nominal scales toward ratio scales, increasingly sophisticated levels of measurement are possible.

4.02.2.1 Scales of Measurement

4.02.2.1.1 Nominal scales

A nominal scale is a qualitative system of categorizing people (or objects, traits, or other variables) or individual observations regarding people typically into mutually exclusive classes or sets. Sex is an example of a nominal scale; one is either male or female. Diagnostic categories such as hyperactivity, learning disabled, aphasia, severely emotionally disturbed, or major depressive disorder represent nominal scaling categories that are not mutually exclusive. Nominal scales provide so little quantitative information about members of categories that some writers prefer to exclude nominal scales from the general rubric of measurement. As Hays (1973) points out, the term measurement typically is reserved for a situation where each individual is assigned a relational number. Because the quantitative relationship among nominal categories is unknown, many common statistical tests cannot be employed with nominal scale data. However, since nominal scales do allow for the classification of an event into a discrete category, many writers (e.g., Nunnally, 1978) do include them as one type of measurement.

4.02.2.1.2 Ordinal scales

Ordinal scales provide considerably more quantitative information regarding an observation than nominal scales. Ordinal scales allow one to rank objects or people according to the amount of a particular attribute displayed. Ordering usually takes the form of the "most" to the "least" amount of the attribute in question. If children in a classroom were weighed and then ranked from heaviest to lightest with the heaviest child assigned the rank of 1, the next heaviest a 2, and so on, until all children had been assigned a number, the resulting measurement would be on an ordinal scale. Although an ordinal scale provides certain quantitative information about each individual, it does not tell how far apart each observation is from the next one. Between adjacent pairs of ranks there may be a different degree of difference. The difference in weight between child 1 and child 2

may be 10 pounds, but the difference between child 2 and child 3 may be one pound or even less. Ordinal scales thus designate relative positions among individuals, an advance over nominal scaling, but are still crude with regard both to describing individuals and to the possible statistical treatments that can be meaningfully applied. Means and standard deviations are usually without meaning when applied to ordinal scales, although the median and mode can be determined and used meaningfully. Age and grade equivalents are examples of common ordinal scales.

4.02.2.1.3 Interval scales

Interval scales afford far more information about observations and can be mathematically manipulated with far greater confidence and precision than nominal or ordinal scales. To have an interval scale of measurement, one must have an ordinal scale on which the difference between any two adjacent points on the scale is equal. Most of the measurement scales and tests used in psychology and education assume an interval scale. Intelligence tests are one good example of an interval scale and can also illustrate the distinction between interval and the highest level of measurement, ratio scales. Although nearly all statistical methods can be applied to measurements on an interval scale, the interval scale has no true zero point, where zero designates total absence of an attribute. If one were to earn an IQ of zero on an intelligence test (by failing to answer a single question correctly), this would not indicate the absence of intelligence, for without intelligence no human could remain alive (it is not possible on most tests of intelligence to earn an IQ of zero even if all test questions are answered incorrectly).

4.02.2.1.4 Ratio scales

Ratio scales possess the attributes of ordinal and interval scales but also have a true zero point—a score of zero indicates the complete absence of the attribute under consideration. Length and width are ratio scales. There are few instances of ratio scales in psychology outside of measurement of simple sensory and motor functions. Ratio scales have useful quantitative features, in particular, as indicated by the name: ratios are meaningful—six feet is twice three feet. Ratios are not meaningful with interval scales. A person with an IQ of 100 cannot be said to be twice as intelligent as a person with an IQ of 50. Fortunately, it is not necessary to have ratio scales to attack the vast majority of problems in psychological assessment.

This discussion of scales of measurement has necessarily been limited to the most basic elements and distinctions among scales. The reader who desires to explore this topic from a technical perspective will find an excellent and extensive mathematical presentation of scales of measurement in Hays (1973).

4.02.2.2 Norms and Reference Groups

To understand the individual's performance as represented by a score on a psychological measurement device, it is necessary, except with certain very specific tests, to evaluate the individual's performance relative to the performance of some preselected group. To know simply that an individual answers 60 out of 100 questions correctly on a history test, and 75 out of 100 questions correctly on a biology test, conveys very little information. On which test did this individual earn the better score? Without knowledge of how a comparable or other relevant group of persons would perform on these tests, the question of which score is better cannot be answered.

Raw scores on a test, such as the number or percentage of correct responses, take on meaning only when evaluated against the performance of a normative or reference group of individuals. For convenience, raw scores are typically converted to a standard or scaled score and then compared against a set of norms. The reference group from which the norms are derived is defined prior to the standardization of the test. Once the appropriate reference population has been defined, a random sample is tested, with each individual tested under as nearly identical procedures as possible. Many factors must be considered when developing norms for test interpretation. Ebel (1972), Angoff (1971), and Petersen, Kolen, and Hoover (1989) have provided especially good discussions of the necessary conditions for appropriate development and use of normative reference data. The following points are taken principally from these three sources, with some elaboration by the present author. Some of these conditions place requirements on the test being normed, some on the psychological trait being measured, and others on the test user.

(i) The psychological trait being assessed must allow the ranking of individuals along a continuum from high to low, that is, it must be amenable to at least ordinal scaling. If a nominal scale was employed, only the presence or absence of the trait would be of interest and relative amounts of the trait could not be determined; norms, under this unusual condition, would be superfluous if not distracting or misleading.

(ii) The content of the test must provide an adequate operational definition of the psychological trait under consideration. With a proper operational definition, other tests can be constructed to measure the same trait and should yield comparable scores for individuals taking both tests.

(iii) The test should assess the same psychological construct throughout the entire range of performance.

(iv) The normative reference group should consist of a large random sample representative of the population on whom the test is to be administered later.

(v) The normative sample of examinees from the population should "have been tested under standard conditions, and . . . take the test as seriously, but no more so, than other(s) to be tested later for whom the norms are needed" (Ebel, 1972, p. 488).

(vi) The population sampled to provide normative data must be appropriate to the test and to the purpose for which the test is to be employed. The latter point is often misinterpreted, especially with regard to evaluation of exceptional children. Many adequately normed psychological tests are inappropriately maligned for failure to include significant numbers of handicapped children in their normative sample. The major intelligence scales designed for use with children (i.e., the various Wechsler scales and the McCarthy Scales of Children's Abilities) have been normed on stratified random samples of children representative of children in the United States. Some authors (e.g., Salvia & Ysseldyke, 1981) criticize tests such as the Wechsler scales as inappropriate for measuring the intellectual level of various categories of children with disabilities because large numbers of these children were not included in the test's standardization sample. Whether this is a valid criticism depends on the purpose to which the test is applied. If knowledge of an emotionally disturbed child's level of intellectual functioning relative to age mates in the United States is desired, comparing the child's performance to that of other similarly emotionally disturbed children, then a reference group of emotionally disturbed children would be appropriate. The latter information is not sought frequently nor has it been shown to be more useful in the diagnosis or development of appropriate intervention strategies. Salvia and Ysseldyke (1981) contend that it would be inappropriate to base predictions of future intellectual or academic performance on test scores for an exceptional child that have been derived through comparison with the larger, normal population's performance. To make predictions, they would first require that the reference group from which

scores are derived be a group of similar socio-cultural background, experience, and handicapping condition. Although this may be an appropriate, if not noble, hypothesis for research, implementation must await empirical verification, especially since it runs counter to traditional psychological practice. Indeed, all interpretations of test scores should be guided principally by empirical evidence. Once norms have been established for a specific reference group, the generalizability of the norms becomes a matter of actuarial research; just as norms based on one group may be inappropriate for use with another group, the norms may also be appropriate and *a priori* acceptance of either hypothesis would be incorrect (Reynolds & Brown, 1984). A large, cumulative body of evidence demonstrates clearly that test scores predict most accurately (and equally well for a variety of subgroups) when based on a large, representative random sample of the population, rather than open highly specific subgroups within a population (e.g., Hunter, Schmidt, & Rauschenberger, 1984; Jensen, 1980; Reynolds, 1982, 1995, in press-a, in press-b).

(vii) Normative data should be provided for as many different groups as it may be useful for an individual to be compared against. Although this may at first glance seem contradictory to the foregoing conclusions, there are instances when it is useful to know how a patient compares to members of other specific subgroups. The more good reference groups available for evaluating a patient's performance on a test, the potentially more useful the test may become.

The normative or reference group most often used to derive scores is the standardization sample, a sample of the target population drawn using a set plan. The best tests, and most publishers and developers of tests, aspire to a standardization sample that is drawn using population proportionate stratified random sampling. This means that samples of people are selected based on subgroups of a larger group to ensure that the population as a whole is represented. In the USA, for example, tests are typically standardized via a sampling plan that stratifies the sample by gender, age, ethnicity, socioeconomic background, region of residence, and community size based on population statistics provided by the US Bureau of the Census. If the Census Bureau data were to indicate, for example, that 1% of the US population consisted of African-American males in the middle range of socioeconomic status residing in urban centers of the south region, then 1% of the standardization sample of the test would be drawn to meet this same set of characteristics.

Once the normative reference group has been obtained and tested, tables of standardized or scaled scores are developed. These tables are based on the responses of the standardization sample and are frequently referred to as norms tables. There are many types of scaled scores or other units of measurement that may be reported in the “norms tables” and just which unit of measurement has been chosen may greatly influence score interpretation.

4.02.3 UNITS OF MEASUREMENT

Raw scores such as number correct are tedious to work with and to interpret properly. Raw scores are thus typically transformed to another unit of measurement. Scaled scores are preferred, but other units such as age and grade equivalents are common. Making raw scores into scaled scores involves creating a set of scores with a predetermined mean and standard deviation that remain constant across some preselected variable such as age.

The mean is simply the sum of the scores obtained by individuals in the standardization sample divided by the number of people in the sample ($\Sigma X_i/N$). In a normal distribution of scores (to be described in the next paragraph), the mean breaks performance on the test into two equal parts, with half of those taking the test scoring above the mean and half scoring below the mean, though the *median* is formally defined as the score point which breaks a distribution into two equal parts; in a normal distribution, the mean and median are the same score.

The standard deviation (SD) is an extremely useful statistic in describing and interpreting a test score. The SD is a measure of the dispersion of scores about the mean. If a test has a mean of 100 and an individual earns a score of 110 on the test, we still have very little information except that the individual performed above average. Once the SD is known, one can determine how far from the mean the score of 110 falls. A score of 110 takes on far different meaning depending on whether the SD of the scores is 5, 15, or 30. The SD is relatively easy to calculate once the mean is known; it is determined by first subtracting each score from the mean, squaring the result, and summing across individuals. This sum of squared deviations from the mean is then divided by the number of persons in the standardization sample. The result is the variance of the test scores; the square root of the variance is the SD.

Once the mean and SD of test scores are known, an individual's standing relative to others on the attribute in question can be

determined. The normal distribution or normal curve is most helpful in making these interpretations. Figure 1 shows the normal curve and its relationship to various standard score systems. A person whose score falls 1 SD above the mean performs at a level exceeding about 84% of the population of test-takers. Two SDs will be above 98% of the group. The relationship is the same in the inverse below the mean. A score of 1 SD below the mean indicates that the individual exceeds only about 16% of the population on the attribute in question. Approximately two-thirds (68%) of the population will score within 1 SD of the mean on any psychological test.

Standard scores such as those shown in Figure 1 (z scores, T scores, etc.) are developed for ease of interpretation. Though standard scores are typically linear transformations of raw scores to a desired scale with a predetermined mean and SD, normalized scaled scores can also be developed. In a linear transformation of test scores to a predetermined mean and SD, equation (1) must be applied to each score:

$$\text{scaled score} = X_{ss} + SD_{ss} \frac{(X_i - X)}{SD_x} \quad (1)$$

where X_i = raw score of any individual i , X = mean of the raw scores, SD_x = standard deviation of the raw scores, SD_{ss} = standard deviation scaled scores are to have, and X_{ss} = mean scaled scores are to have.

Virtually all tests designed for use with children along with most adult tests standardize scores and then normalize them within age groups so that a scaled score at one age has the same meaning and percentile rank at all other ages. Thus a person age 10 who earns a scaled score of 105 on the test has the same percentile rank within his or her age group as a 12-year-old with the same score has in his or her age group. That is, the score of 105 will fall at the same point on the normal curve in each case.

Not all scores have this property. Grade and age equivalents are very popular types of scores that are much abused because they are assumed to have scaled score properties when in fact they represent only an ordinal scale. Grade equivalents ignore the dispersion of scores about the mean although the dispersion changes from age to age and grade to grade. Under no circumstances do such equivalent scores qualify as standard scores. Consider the calculation of a grade equivalent. When a test is administered to a group of children, the mean raw score is calculated at each grade level and this mean raw score then is called the “grade equivalent” score for a raw score of that magnitude. If the mean raw score for beginning fourth graders (grade 4.0) on a reading test is 37, then any person

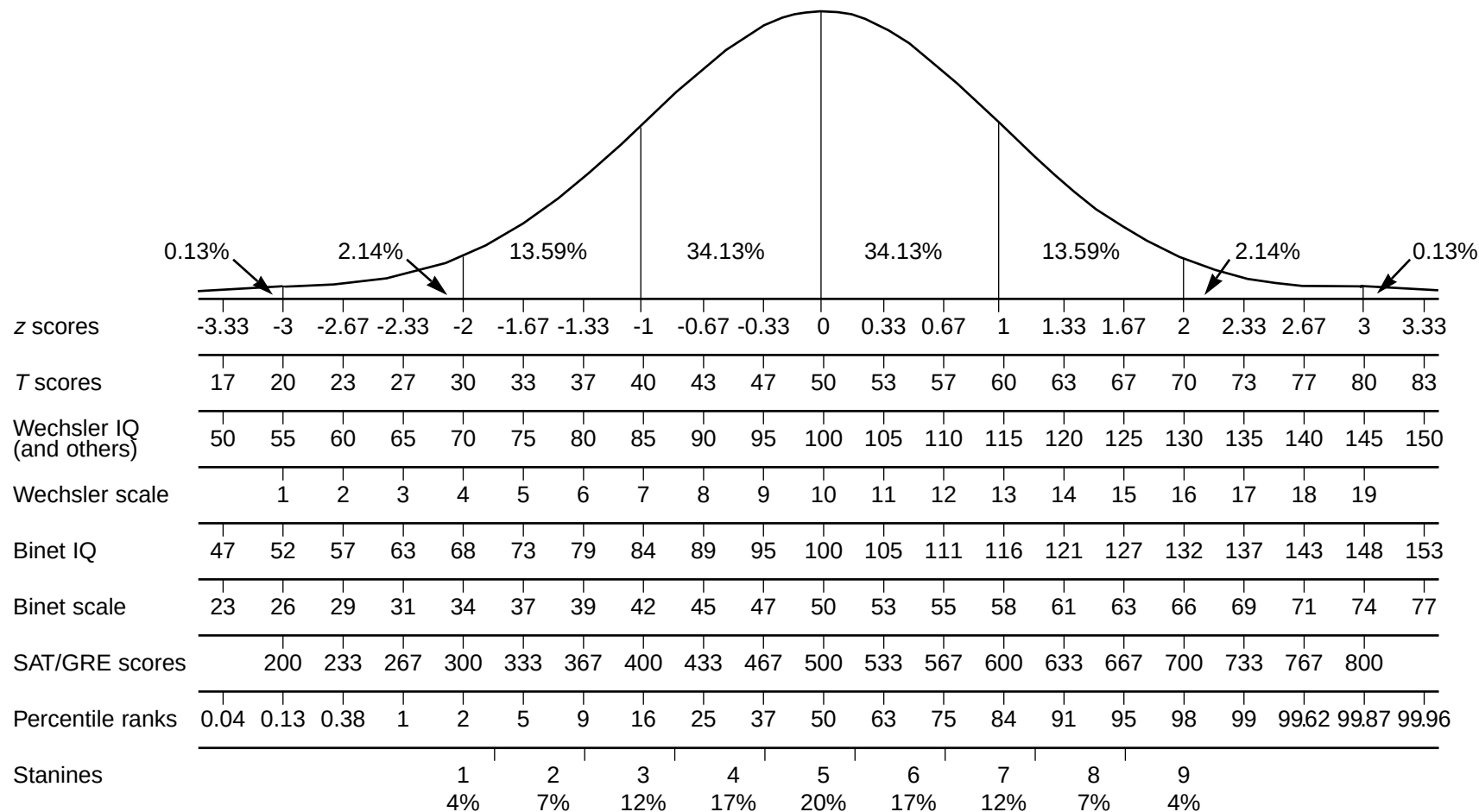


Figure 1 Relationships among the normal curve, relative standing expressed in percentiles, and various systems of derived scores.

earning a score of 37 on the test is assigned a grade equivalent score of 4.0 regardless of the person's age. If the mean raw score of fifth graders (grade 5.0) is 38, then a score of 38 would receive a grade equivalent of 5.0. A raw score of 37 could represent a grade equivalent of 4.0, 38 could be 5.0, 39 could be 5.1, 40 be 5.3, and 41, 6.0. Thus, differences of one raw score point can cause dramatic differences in the grade equivalents received, and the differences will be inconsistent across grades with regard to the magnitude of the difference in grade equivalents produced by constant changes in raw scores.

Table 1 illustrates the problems of using grade equivalents to evaluate a patient's academic standing relative to his or her peers. Frequently in both research and clinical practice, children of normal intellectual capacity are diagnosed as learning disabled through the use of grade equivalents such as "two years below grade level for age" on a test of academic attainment. The use of this criterion for diagnosing learning disabilities or other academic disorders is clearly inappropriate (Reynolds, 1981a, 1985). As seen in Table 1, a child with a grade equivalent score in reading two years below the appropriate grade placement for age may or may not have a reading problem. At some ages this is within the average range, whereas at others a severe reading problem may be indicated.

Grade equivalents tend to become standards of performance as well, which they clearly are not. Contrary to popular belief, grade equivalent scores on a test do not indicate what level of reading text a child should be using. Grade equivalent scores on tests simply do not have a one-to-one correspondence with reading series placement or the various formulas for determining readability levels.

Grade equivalents are also inappropriate for use in any sort of discrepancy analysis of an individual's test performance, diagnosis of a learning disability or developmental disorder, or for use in many statistical procedures for the following reasons (Reynolds, 1981a).

(i) The growth curve between age and achievement in basic academic subjects flattens at upper grade levels. This can be seen in Table 1 where there is very little change in standard score values corresponding to two years below grade level for age after about grade 7 or 8. In fact, grade equivalents have almost no meaning at this level since reading instruction typically stops by high school. Consider the following analogy with height as an age equivalent. Height can be expressed in age equivalents just as reading can be expressed as grade equivalents. It might be helpful to describe a tall first grader as having the height of an 8½ year old,

but what happens to the 5 feet, 10 inch tall 14-year-old female since at no age does the mean height of females equal 5 feet, 10 inches? Since the average reading level in the population changes very little after junior high school, grade equivalents at these ages become virtually nonsensical with large fluctuations resulting from a raw score difference of two or three points on a 100-item test.

(ii) Grade equivalents assume that the rate of learning is constant throughout the school year and that there is no gain or loss during summer vacation.

(iii) Grade equivalents involve an excess of extrapolation, especially at the upper and lower ends of the scale. However, since tests are not administered during every month of the school year, scores between the testing intervals (often a full year) must be interpolated on the assumption of constant growth rates. Interpolations between sometimes extrapolated values on an assumption of constant growth rates is a somewhat ludicrous activity.

(iv) Different academic subjects are acquired at different rates and the variation in performance varies across content areas so that "two years below grade level for age" may be a much more serious deficiency in math than in reading comprehension.

(v) Grade equivalents exaggerate small differences in performance between individuals and for a single individual across tests. Some test authors even provide a caution on record forms that standard scores only, and not grade equivalents, should be used for comparisons.

Age equivalents have many of the same problems. The standard deviation of age equivalents varies substantially across tests, subsets, abilities, or skills assessed, and exist on an ordinal, not interval scale. It is inappropriate to add, subtract, multiply, or divide age or grade equivalents or any other form of ordinal score. Nevertheless, the use of such equivalent scores in *ipsative* analysis of test performance remains a common mistake in clinical, educational, and neuropsychological assessment.

The principal advantage of standardized or scaled scores lies in the comparability of score interpretation across age. By standard scores of course, I refer to scores scaled to a constant mean and SD such as the Wechsler Deviation IQ and not to ratio IQ types of scales employed by the early Binet and original Slosson Intelligence Test, which give the false appearance of being scaled scores. Ratio IQs or other types of quotients have many of the same problems as grade equivalents and should be avoided for many of these same reasons. Standard scores of the deviation IQ type have the same percentile rank across age since they

Table 1 Standard scores and percentile ranks corresponding to performance “two years below grade level for age” on three reading tests.

Grade placement	Two years below placement	Wide range achievement test		Woodcock Reading Mastery Test ^a		Stanford Diagnostic Reading Test ^a	
		SS ^b	%R ^c	SS	%R	SS	%R
2.5	K.5	67	1				
3.5	1.5	69	2	64	1	64	1
4.5	2.5	73	4	77	6	64	1
5.5	3.5	84	14	85	16	77	6
6.5	4.5	88	21	91	27	91	27
7.5	5.5	86	18	94	34	92	30
8.5	6.5	87	19	94	34	93	32
9.5	7.5	90	25	96	39	95	34
10.5	8.5	85	16	95	37	95	37
11.5	9.5	85	16	95	37	92	30

^a Total test. ^b All standard scores in this table have been converted for ease of comparison to a common scale having a mean of 100 and an SD of 15. ^c Percentile rank.

Source: Adapted from Reynolds (1981a).

are based not only on the mean but the variability in scores about the mean at each age level. For example, a score that falls two-thirds of a SD below the mean has a percentile rank of 25 at every age. A score falling two-thirds of a grade level below the average grade level or an age equivalent six months below chronological age have different percentile ranks at every age.

Standard scores are more accurate and precise. When constructing tables for the conversion of raw scores into standard scores, interpolation of scores to arrive at an exact score point is typically not necessary, whereas the opposite is true of age and grade equivalents. Extrapolation is also typically not necessary for scores within 3 SDs of the mean, which accounts for more than 99% of all scores encountered.

Scaled scores can be set to any desired mean and standard deviation, with the fancy of the test author frequently the sole determining factor. Fortunately, a few scales can account for the vast majority of standardized tests in psychology and education. Table 2 illustrates the relationship between various scaled score systems. If reference groups are comparable, Table 2 can also be used to equate scores across tests to aid in the comparison of a patient's performance on tests of different attributes, provided normalized scores are provided.

What has been said thus far about scaled scores and their equivalency applies primarily to scores that have been forced to take the shape of the Gaussian or bell curve. When test-score distributions derived from a standardization sample are examined, the scores frequently deviate significantly from normal. Often, test

developers will then transform scores, using one of a variety of statistical methods (e.g., see Lord & Novick, 1968, for a mathematical review and explication), to take a normal distribution. Despite what is often taught in early courses in psychological statistics and measurement, this is not always appropriate. It is commonplace to read that psychological variables, like most others, are normally distributed within the population; many are. Variables such as intelligence, memory skill, and academic achievement will closely approximate the normal distribution when well measured. However, many psychological variables, especially behavioral ones such as aggression, attention, and hyperactivity, deviate substantially from the normal curve within the population of humans.

When a score distribution then deviates from normality, the test developer is faced with the decision of whether to create normalized scores via some transformation or to allow the distribution to retain its shape with perhaps some smoothing to remove irregularities due to sampling error. In the later case, a linear transformation of scores is most likely to be chosen.

To make this determination, the test developer must ascertain whether the underlying construct measured by the test is normally distributed or not and whether the extant sample is adequate to estimate the distribution, whatever its shape. For applied, clinical devices, the purpose of score transformations that result in normalization of the distribution is to correct for sampling error and presumes that the underlying construct is, in fact, normally or near normally distributed. Normalization of the score distribution then produces a more

Table 2 Conversion of standard scores based on several scales to a commonly expressed metric.^a

Scales									Percentile rank
$X = 0$ $SD = 1$	$X = 10$ $SD = 3$	$X = 36$ $SD = 6$	$X = 50$ $SD = 10$	$X = 50$ $SD = 15$	$X = 100$ $SD = 15$	$X = 100$ $SD = 16$	$X = 100$ $SD = 20$	$X = 500$ $SD = 100$	
2.6	18	52	76	89	139	142	152	760	> 99
2.4	17	51	74	86	136	138	148	740	99
2.2	17	49	72	83	133	135	144	720	99
2.0	16	48	70	80	130	132	140	700	98
1.8	15	47	68	77	127	129	136	680	96
1.6	15	46	66	74	124	126	132	660	95
1.4	14	44	64	71	121	122	128	640	92
1.2	14	43	62	68	118	119	124	620	88
1.0	13	42	60	65	115	116	120	600	84
0.8	12	41	58	62	112	113	116	580	79
0.6	12	40	56	59	109	110	112	560	73
0.4	11	38	54	56	106	106	108	540	66
0.2	11	37	52	53	103	103	104	520	58
0.0	10	36	50	50	100	100	100	500	50
-0.2	9	35	48	47	97	97	96	480	42
-0.4	9	34	46	44	94	94	92	460	34
-0.6	8	33	44	41	91	90	88	440	27
-0.8	8	31	42	38	88	87	84	420	21
-1.0	7	30	40	35	85	84	80	400	16
-1.2	6	29	38	32	82	81	76	380	12
-1.4	6	28	36	29	79	78	72	360	8
-1.6	5	26	34	26	76	74	68	340	5
-1.8	5	25	32	23	73	71	64	320	4
-2.0	4	24	30	20	70	68	60	300	2
-2.2	3	23	28	17	67	65	56	280	1
-2.4	3	21	26	14	64	62	52	260	1
-2.6	2	20	24	13	61	58	48	240	1

^a X = mean; SD = standard deviation.

accurate rendition of the population distribution and improves the utility of the standardized scaled scores provided. If the population distribution of the construct in question is not normal, for example, aggressive behavior (see Reynolds & Kamphaus, 1992), then a different form of transformation, typically linear, is required to be accurate. This decision affects how clinicians best interpret the ultimately scaled scores.

If score distributions have been normalized for a battery of tests or subtests of a common test, for example, the Wechsler scales, the same scaled score on any part-test will have the same percentile rank. On the Wechsler Intelligence Scale for Children-III (WISC-III; Wechsler, 1992), for example, a subtest scaled score of 13 is 1 SD above the mean and, for all 13 subtests of the WISC-III, will have a percentile rank of approximately 86. If the scores had not been transformed through the nonlinear methods necessary to approximate a normal distribution, this would not be true. For a linear transformation, a scaled score of 13 could still be 1 SD

above the mean on all of the subtests but the percentile rank could vary, and could vary substantially the more the underlying distribution deviates from that of the normal curve. It is thus important for clinicians to review test manuals and ascertain the methods of scaling that have been applied to the raw score distributions. This becomes increasingly important as scores are to be compared across different tests or batteries of tests. This effect is magnified as the distance from the mean increases.

4.02.4 ACCURACY OF TEST SCORES

4.02.4.1 True Score Theory

When evaluating test scores, it is also necessary to know just how accurately the score reflects the individual's true score on the test. Tests typically do not ask every possible question that could be asked or evaluate every possible relevant behavior. Rather a domain of possible questions or test items is defined and a

sample taken to form the test. Whenever less than the total number of possible behaviors within a domain is sampled, sampling error occurs. Psychological and educational tests are thus destined to be less than perfect in their accuracy. Certainly, psychological tests contain errors produced from a variety of other sources, most of which are situational. Error resulting from domain sampling is the largest contributor to the degree of error in a test score, however (Feldt & Brennan, 1989; Nunnally, 1978), and is the type of error about which measurement theory has the greatest concern. Fortunately, this type of error is also the easiest and most accurately estimated.

Error caused by domain sampling is determined from an analysis of the degree of homogeneity of the items in the test, that is, how well the various items correlate with one another and with an individual's true standing on the trait being assessed. The relative accuracy of a test is represented by a reliability coefficient symbolized as r_{xx} . Since it is based on the homogeneity or consistency of the individual items of a test and no outside criteria or information are necessary for its calculation, r_{xx} is frequently referred to as internal consistency reliability or as an estimate of item homogeneity. Error caused by domain sampling is also sometimes estimated by determining the correlation between two parallel forms of a test (forms of a test that are designed to measure the same variable with items sampled from the same item domain and believed to be equivalent). The correlation between the two equivalent or alternate forms is then taken as the reliability estimate and is usually symbolized as r_{xx} , r_{ab} , or r_{xy} (although r_{xy} is generally used to represent a validity coefficient).

Split-half reliability estimates can also be determined on any specific test as a measure of internal consistency. Split-half reliability is typically determined by correlating each person's score on the one-half of the items with his or her score on the other half of the test with a correction for the original length of the test, since length will affect reliability. Predetermined or planned split-half comparisons such as correlating scores on odd numbered items with scores on the even numbered items may take advantage of chance or other factors resulting in spuriously high estimates of reliability. A reliability coefficient called alpha is a better method for estimating reliability since it is the mean of all possible split-half comparisons, thus expunging any sampling error resulting from the method of dividing the test for the purposes of calculating a correlation between each half.

As noted earlier, a number of techniques exist for estimating reliability. Throughout this

chapter, reliability has been referred to as estimated. This is because the absolute or "true" reliability of a psychological test can never be determined. Alpha and all other methods of determining reliability are, however, considered to be lower bound estimates of the true reliability of the test. One can be certain that the reliability of a test is at least as high as the calculated estimate and possibly even higher.

Once the reliability of a test has been estimated, it is possible to calculate a sometimes more useful statistic known as the standard error of measurement. Since there is always some error involved in the score a person obtains on a psychological test, the obtained score (X_i) does not truly represent the individual's standing with regard to the trait in question. Obtained scores estimate an individual's true score on the test (the score that would be obtained if there was no error involved in the measurement). Since this is not possible, the true score (X_∞) is defined as the mean score of an individual if administered an infinite number of equivalent forms of a test and there were no practice effects or other intervening factors. The standard error of measurement (S_{em}) is the SD of the individual's distribution of scores about his or her true score. To determine the S_{em} it is necessary to know only the SD and the reliability (preferably an internal consistency estimate) of the test in question. The calculation of X_∞ and S_{em} are only estimates, however, since the conditions for determining a true score never actually exist.

Since the distribution of obtained scores about the true score is considered to be normal, one can establish a degree of confidence in test results by banding the estimated true score by a specified number of S_{ems} . A table of values associated with the normal curve (pictured in Figure 1) quickly tells us how many S_{ems} are necessary for a given level of confidence. In a normal distribution, about 68% of all scores fall within 1 SD of the mean, and about 95% of all scores fall within 2 SDs of the mean. Therefore, if one wanted to be 68% certain that a range of scores contained a person's true score, X_∞ would be banded by $\pm 1 S_{em}$. To be 95% certain that a range of scores contained the true score, a range of $X_\infty \pm 2 S_{ems}$ would be necessary.

When evaluating a test or performance on a test, it is important to ascertain just what type of reliability estimate is being reported. S_{ems} should typically be calculated from an internal consistency estimate. Comparisons of reliability estimates across tests should be based on the same type of estimate. For example, one should not compare the reliability of two tests based on alternate form correlations for one test and estimation of the alpha coefficient for the other.

Test–retest correlations, also frequently referred to as reliability coefficients, should not be confused with measures of the accuracy or precision of a test at a given point in time.

Test–retest “reliability” is one of the most often confused concepts of psychometric theory. Even Anastasi (1976), in introducing reliability, refers to reliability as a measure of the degree to which a person would obtain the same score if tested again at a later time. In the earlier stages of development of psychology when traits were considered unchanging, test–retest reliability was properly considered to be a characteristic of the test and indeed was believed to be an indication of the degree to which a person would obtain the same score if tested again. Test–retest reliability speaks principally of the stability of the trait being measured and has little to do with the accuracy or precision of measurement unless the psychological construct in question is considered to be totally unchangeable. Given that traits such as anxiety and even intelligence do in fact change over time and that testing from one time to the next is positively correlated, it is still possible to use the test–retest correlation to determine estimates of what score a person would obtain upon retesting. Internal consistency estimates, however, should not be interpreted in such a manner. When psychological constructs are not highly labile and believed to change only over long periods of time, test–retest correlations may be considered to reflect the accuracy of a test if the two testings occur at close points in time during which the trait under consideration is believed to be stable.

4.02.4.2 Generalizability Theory

Generalizability theory is an extension of true score theory (also known as classical test theory) that is achieved principally through use of analysis of variance (ANOVA) procedures. Often, more than one type of error is acting on a reliability coefficient. For example, in true score theory, errors due to domain sampling (e.g., not asking about every possible symptom of depression), errors due to faulty administration, scoring errors by the examiner, and errors associated with time sampling may all act to lower the average interitem correlation, which will reduce the internal consistency reliability of the test score. Under true score theory, it is impossible to partial the relative contributions, that is, to determine how much error is contributed by each subset of error to the total amount of unreliability. Even test–retest or stability coefficients are confounded by internal consisting errors. The maximum r_{12} is equal to the square root of r_{xx} or $\max r^{12} = (r_{xx})^{1/2}$.

Generalizability theory takes advantage of the capabilities of ANOVA in partitioning variance components to develop a model of unreliability (as opposed to concentrating on statistical significance). Through ANOVA, generalizability theory is able to partition the error variance of a set of scores into the components listed above, such as domain sampling error and the like, along with some additional components not considered in true score theory. Generalizability theory is no more difficult mathematically than true score theory. Generalizability theory is surprisingly absent from the measurement repertoire of most clinicians but is becoming increasingly popular among measurement scientists. However, the understanding and application of generalizability theory does require an understanding of methods and designs for partitioning variance components in ANOVA, a skill that is perhaps on the decline in clinical training programs in favor of statistical methods more aligned with structural equation modeling.

The basic foundations of generalizability theory can be found in Cronbach, Rajaratnam, and Gleser (1963). A current, detailed explanation appears in Feldt and Brennan (1989) along with the necessary mathematical models necessary to apply generalizability theory to the concept of error in test scores of groups and individuals.

4.02.5 VALIDITY

Reliability refers to the precision or accuracy of test scores. Validity refers to the appropriateness of the interpretations of test scores and not to the test or the score itself. “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and the appropriateness of inferences and actions based on test scores or other modes of assessment” (Messick, 1989, p. 13). As is reliability, validity is a matter of degree and not an all or none concept. Reliability will, however, enter into evaluation of the validity of an inference drawn from a test score. Reliability is a necessary but insufficient condition for validity. As reliability approaches zero, the amount of random error in a test score increases. The greater the relative proportion of random error present, the less confidence one can have in any interpretation of a score since, by definition, random error is unrelated to anything meaningful. Validation is not static but is an ongoing process, not just of the corroboration of a particular meaning, but for the development of sounder and better interpretations of observations that are expressed as scores on a psychological test.

Although it is often done as a matter of convenience or as simple short hand, it should be obvious by now that it is not correct technically to refer to the validity of a test. Validity is a characteristic of the interpretation given to performance on a test. It makes no sense, for example, to ask a question such as "Is this Wechsler scale a valid test?" Rather, one might pose the superior question "Is the interpretation of performance on this Wechsler scale as reflecting intelligence or intellectual level valid?" This is more than a game of semantics as such subtle differences in language affect the way we think about our methods and our devices. The difference in language and its implications are considered powerful enough that *Educational and Psychological Measurement*, one of the oldest and most respected journals in psychometrics, founded originally by Frederic Kuder, no longer allows authors in its pages to refer to the validity of a test or the reliability of a test. Reviewers for this journal are asked routinely to screen manuscripts for improper or imprecise use of such terminology.

Just as reliability may take on a number of variations, so may validity. Quite a bit of divergent nomenclature has been applied to validity. Messick (1980) identified 17 "different" types of validity that are referred to in the technical literature! Traditionally, validity has been broken into three major categories: content, construct, and predictive or criterion-related validity. These are the three types of validity distinguished and discussed in the joint *Standards for Educational and Psychological Tests* (American Psychological Association, 1985). Construct validity cuts across all categories, and criterion-related validity is definitely a question of the relationship of test performance to other methods of evaluating behavior.

Content validity is determined by how well the test items and their specific content sample the set of behaviors or subject matter area about which inferences are to be drawn on the basis of the test scores. Criterion-related or predictive validity refers to either comparisons of test scores with performance on accepted criteria of the construct in question taken in close temporal relationship to the test or the level of prediction of performance at some future time. Criterion-related validity is determined by the degree of correspondence between the test score and the individual's performance on the criterion. If the correlation between these two variables is high, no further evidence may be considered necessary (Nunnally, 1978). Here, reliability has a direct, and known, limiting effect on validity. A correlation between a predictor (x) and a criterion (y), a validity

coefficient, typically expressed as r_{xy} , is restricted in magnitude. Its maximum true value is equal to the square root of the product of the internal consistency reliability coefficients of the scores being compared: $r_{xy} \max = (r_{xx} r_{yy})^{1/2}$.

Construct validity of the interpretations given to psychological tests is one of the most complex issues facing the psychometrician and permeates all aspects of test development and test use. Psychology for the most part deals with intangible constructs. Intelligence is one of the most intensely studied constructs in the field of psychology, yet it cannot be directly observed or evaluated. Intelligence can only be inferred from the observation and quantification of what has been agreed upon as "intelligent" behavior. Personality variables such as dependence, anxiety, need achievement, mania, and on through the seemingly endless list of personality traits that psychologists have "identified" also cannot be observed directly. Their existence is only inferred from the observation of behavior. Construct validity thus involves considerable inference on the part of the test developer and the researcher; construct validity is evaluated by investigating just what psychological properties a test measures.

Prior to being used for other than research purposes, interpretations given to a test must be shown clearly to demonstrate an acceptable level of validity. For use with various categories of psychopathology, validation with normally functioning individuals should be considered insufficient. The validity of an interpretation needs to be demonstrated for each group with whom it is used. This can be a long and laborious process but is nevertheless a necessary one. There are many subtle characteristics of various classes of exceptional children, for example, that may cause an otherwise appropriate interpretation of a test to lack validity with special groups (e.g., see Newland, 1980).

As has been noted by Cronbach (1971) and others, the term "test validation" can cause some confusion. In thinking about and evaluating validity, we must always keep in mind that one does not ever actually validate a test but only the interpretation that is given to the score on the test. Any test may have many applications and a test with originally a singular purpose may prove promising for other applications. Each application of a test or interpretation of a test score must undergo validation. Whenever hearing or reading that a test has been validated, we need to know for what purpose it has been validated, and what interpretations of scores from the instrument in question have been shown empirically to be justifiable and accurate.

4.02.6 THE ASSESSMENT PROCESS

As noted at the opening of this chapter, assessment is an involved, comprehensive process of deriving meaning from test scores to achieve a broad but detailed description and understanding of the individual. The description here of assessment as a process is important. Assessment, properly carried out, is not a static collection of information, but an ongoing dynamic synthesis and evaluation of data, reliably obtained, from multiple sources relevant to the current, and possibly future, status of the individual. Assessment is open ended—new information can occur daily that can properly alter one's perception of the ecological validity of prior impressions and recommendations.

Crucial to the assessment process, and far too frequently neglected or overlooked, is follow-up evaluation that should occur after more formal diagnostic assessments have been made and habilitative recommendations implemented. There are no absolutes in psychological and educational testing; no profile of assessment information is inexorably linked with a single method of treatment, remediation, or intervention that will always be successful. Currently, the opposite is the case; the search for the aptitude $\times$ treatment interaction is nearly as elusive as that for the neural engram. The follow-up component of the assessment process is crucial to the fine-tuning of existing intervention procedures and in some cases more massive overhauling of an intervention.

Psychological and educational testing and assessment are far from exact, just as are the clinical assessment procedures of medicine and related specialties. When used in diagnosis, assessment allows one simply to narrow the number of disorders under serious consideration. Similarly, when used in the search for an appropriate method of habilitation for a handicapped youngster, the assessment process allows the psychologist to narrow the number of strategies (i.e., hypotheses) from which to choose one that is believed to be most effective. There are no guarantees that the first strategy adopted will be the most effective program of treatment (or be effective at all for that matter). Kaufman (1994) described the proper attitude of the psychologist involved in assessment to be that of a "detective" who evaluates, synthesizes, and integrates data gleaned from the assessment process with knowledge of psychological theories of development and the psychology of individual differences (also see Reynolds, 1981b; Reynolds & Clark, 1982). As described here, the assessment process is a major component in psychological problem-solving. Individuals are not randomly selected for an expensive, time-

consuming psychological evaluation. They are referred to a psychologist for some more or less specific reason; a problem of some kind exists. The assessment process then plays a major role in accurately identifying and describing the problem, suggesting solutions, and properly carried through, provides ideas for modifying the initially proposed interventions.

It is necessary in the assessment process to entertain and evaluate information from a variety of sources if the assessment is to be ecologically valid. Each situation will dictate the relevance and appropriate weighting of each piece of information. Age and physical condition are two obvious factors that influence the gathering of information regarding child and adult patients. Palmer (1980), Newland (1980), Salvia and Ysseldyke (1981), and Sattler (1988) have discussed factors to be included in the assessment process when evaluating exceptional children in the schools. The following are generally accepted to be important aspects of assessment: medical condition, sensory and motor skills, school performance and behavior (e.g., group achievement tests, grades, teacher checklists), individual intelligence test scores, special aptitude and achievement test performance, affective characteristics (e.g., personality tests), teacher reports on behavior and peer interaction, the child-school interaction, characteristics of the classroom, parent reports on behavior, the social and cultural milieu of the home, and the child's developmental history. Each of these factors takes on more or less importance for individual patients. With adult patients, many of the same types of information will be relevant with a conceptual shift toward adulthood (Anastasi & Urbina, 1997). The patient's vocational functioning and relationships including parents, spouse, and children will all need to be considered when designing the assessment and later when interpreting the results. More specialized types of knowledge may be required for any given case. For example, in certain genetically-based disorders, a complete family history may be necessary to achieve a good understanding of the nature of the patient's difficulty.

Numerous methods of psychological testing can be used in the assessment process. Each will have its own strengths and weaknesses. There are frequent debates in the psychological literature over the relative merits of one category of assessment over another, with some respondents carrying on with nearly religious fervor. However, these arguments can be resolved quickly by recalling that tests are tools of assessment and most certainly not an end in themselves. Different methods and techniques of testing are best seen and used as complementary in

assessment, which is a problem-solving process requiring much information. With these admonitions in mind, it is time to turn to a discussion of various methods of testing and their role in the assessment process.

4.02.7 MODELS AND METHODS OF ASSESSMENT

A variety of assessment methods are available for evaluating adults and exceptional children. Some of these methods grew directly from specific schools of psychological thought, such as the psychoanalytic view of Freud (projective assessment techniques) or the behavioral schools of Watson, Skinner, and Bandura (applied behavior analysis). Other methods have grown out of controversies in and between existing academic disciplines such as personality theory and social psychology. New and refined methods have come about with new developments in medicine and related fields, whereas other new testing methods stem from advances in the theory and technology of the science of psychological measurement. Unfortunately, still other new techniques stem from psychological and educational faddism with little basis in psychological theory and little if any empirical basis. Any attempt to group tests by characteristics such as norm-referenced vs. criterion-referenced, traditional vs. behavioral, maximum vs. typical performance, and so on, is doomed to criticism. As will be seen in the pages that follow, the demarcations between assessment methods and models are not so clear as many would contend. In many cases, the greatest distinctions lie in the philosophical orientation and intent of the user. As one prominent example, many “behavioral” assessment techniques are as bound by norms and other traditional psychometric concepts as are traditional intelligence tests (Cone, 1977). Even trait measures of personality end up being labeled by some as behavioral assessment devices (e.g., Barrios, Hartmann, & Shigetomi, 1981). The division of models and methods of assessment to follow is based in some part on convenience and clarity of discussions but also with an eye toward maintaining the most important conceptual distinctions among these assessment methods.

4.02.7.1 Traditional Norm-referenced Assessment

4.02.7.1.1 *Intelligence, achievement, and special abilities*

These assessment techniques have been grouped together primarily because of their similarity of content and, in some cases, their

similarity of purpose. There are, however, some basic distinctions among these measures. Intelligence tests tend to be broad in terms of content; items sample a variety of behaviors that are considered to be intellectual in nature. Intelligence tests are used both to evaluate the current intellectual status of the individual and to predict future behavior on intellectually demanding tasks and to help achieve a better understanding of past behavior and performance in an intellectual setting. Achievement tests measure relatively narrowly defined *content*, sampled from a specific subject matter domain that typically has been the focus of purposeful study and learning by the population for whom the test is intended. Intelligence tests by contrast are oriented more toward testing intellectual processes and use items that are more related to incidental learning and not as likely to have been specifically studied as are achievement test items. Tests of special abilities, such as memory, mechanical aptitude, and auditory perception, are narrow in scope as are achievement tests but focus on process rather than content. The same test question may appear on an intelligence, achievement, or special ability test, however, and closely related questions frequently do. Tests of intelligence and special abilities also focus more on the application of previously acquired knowledge, whereas achievement tests focus on testing just what knowledge has been acquired. One should not focus on single items; it is the collection of items and the use and evaluation of the individual's score on the test that are the differentiating factors.

(i) *Intelligence tests*

Intelligence tests are among the oldest devices in the psychometric arsenal of the psychologist and are likely the most frequently used category of tests in the evaluation of exceptional children, especially in the cases of mental retardation, learning disabilities, and intellectual giftedness. Intelligence and aptitude tests are used frequently in adult assessment as well and are essential diagnostic tools when examining for the various dementias. They are used with adults in predicting a variety of other cognitive disorders and in the vocational arena. Since the translation and modification of Alfred Binet's intelligence test for French schoolchildren was introduced in the United States by Lewis Terman (of Stanford University, hence the Stanford–Binet Intelligence Scale), a substantial proliferation of such tests has occurred. Many of these tests measure very limited aspects of intelligence (e.g., Peabody Picture Vocabulary Test, Columbia Mental Maturity Scale, Ammons and Ammons Quick Test), whereas others

give a much broader view of a person's intellectual skills, measuring general intelligence as well as more specific cognitive skills (e.g., the various Wechsler scales). Unfortunately, while intelligence is a hypothetical psychological construct, most intelligence tests were developed from a primarily empirical basis, with little if any attention given to theories of the human intellect. Empiricism is of major importance in all aspects of psychology, especially psychological testing, but is insufficient in itself. It is important to have a good theory underlying the assessment of any theoretical construct such as intelligence.

Intelligence tests in use today are for the most part individually administered (i.e., a psychologist administers the test to an individual in a closed setting with no other individuals present). For a long time, group intelligence tests were used throughout the public schools and in the military. Group tests of intelligence are used more sparingly today because of their many abuses in the past and the limited amount of information they offer about the individual. There is little of utility to the schools, for example, that can be gleaned from a group intelligence test that cannot be obtained better from group achievement tests. Individual intelligence tests are far more expensive to use but offer considerably more and better information. Much of the additional information, however, comes from having a highly trained observer (the psychologist) interacting with the person for more than an hour in a quite structured setting, with a variety of tasks of varying levels of difficulty. The most widely used individually administered intelligence scales today are the Wechsler scales, the Kaufman scales, and the Stanford-Binet Intelligence Scale (Fourth Edition). Though the oldest and best known of intelligence tests, the Binet has lost much of its popularity and is now a distant third.

Intelligence testing, which can be very useful in clinical and vocational settings, is also a controversial activity, especially with regard to the diagnosis of mild mental retardation among minority cultures in the United States. Used with care and compassion, as a tool toward understanding, such tests can prove invaluable. Used recklessly and with rigidity, they can cause irreparable harm. Extensive technical training is required to master properly the administration of an individual intelligence test (or any individual test for that matter). Even greater sensitivity and training are required to interpret these powerful and controversial devices. Extensive knowledge of statistics, measurement theory, and the existing research literature concerning testing is a prerequisite to using

intelligence tests. To use them well requires mastery of the broader field of psychology, especially differential psychology, the psychological science that focuses on the psychological study and analysis of human individual differences and theories of cognitive development. Extensive discussions of the clinical evaluation of intelligence can be found in Kaufman (1990, 1994) and Kaufman and Reynolds (1983).

(ii) *Achievement tests*

Various types of achievement tests are used throughout the public schools with regular classroom and exceptional children. Most achievement tests are group tests administered with some regularity to all students in a school or system. Some of the more prominent group tests include the Iowa Test of Basic Skills, the Metropolitan Achievement Test, the Stanford Achievement Test, and the California Achievement Test. These batteries of achievement tests typically do not report an overall index of achievement but rather report separately on achievement in such academic areas as English grammar and punctuation, spelling, map reading, mathematical calculations, reading comprehension, social studies, and general science. The tests change every few grade levels to accommodate changes in curriculum emphasis. Group achievement tests provide schools with information concerning how their children are achieving in these various subject areas relative to other school systems throughout the country and relative to other schools in the same district. They also provide information about the progress of individual children and can serve as good screening measures in attempting to identify children at the upper and lower ends of the achievement continuum. Group administered achievement tests help in achieving a good understanding of the academic performance of these individuals but do not provide sufficiently detailed or sensitive information on which to base major decisions. When decision-making is called for or an in-depth understanding of a child's academic needs is required, individual testing is needed.

Psychologists use achievement measures with adult clients as well. With the elderly, acquired academic skills tend to be well preserved in the early stages of most dementias and provide a good baseline of promorbid skills. Academic skills can also be important in recommending job placements, as a component of child custody evaluations, in rehabilitation planning, and in the diagnosis of adult learning disorders and adult forms of attention deficit hyperactivity disorder.

(iii) Tests of special abilities

These are specialized methods for assessing thin slices of the spectrum of abilities for any single individual. These measures can be helpful in further narrowing the field of hypotheses about an individual's learning or behavior difficulties when used in conjunction with intelligence, achievement, and personality measures. The number of special abilities that can be assessed is quite large. Some examples of these abilities include visual-motor integration skills, auditory perception, visual closure, figure-ground distinction, oral expression, tactile form recognition, and psychomotor speed. While these measures can be useful, depending on the questions to be answered, one must be particularly careful in choosing an appropriate, valid, and reliable measure of a special ability. The use and demand for these tests are significantly less than that for the most popular individual intelligence tests and widely used achievement tests. This in turn places some economic constraints on development and standardization procedures, which are very costly enterprises when properly conducted. One should always be wary of the "quick and dirty" entry into the ability testing market. There are some very good tests of special abilities available, although special caution is needed. For example, simply because an ability is named in the test title is no guarantee that the test measures that particular ability. As with all tests, what is actually measured by any collection of test items is a matter for empirical investigation and is subject to the process of validation.

To summarize, norm-referenced tests of intelligence, achievement, and special abilities provide potentially important information in the assessment process. Yet each supplies only a piece of the required data. Equally important are observations of how the patient behaves during testing and in other settings, and performance on other measures.

4.02.7.2 Norm-referenced, Objective Personality Measures

Whereas tests of aptitude and achievement can be described as maximum performance measures, tests of personality can be described as typical performance measures. When taking a personality test, one is normally asked to respond according to one's typical actions and attitudes and not in a manner that would present the "best" possible performance (i.e., most socially desirable). The "faking" or deliberate distortion of responses is certainly possible, to a greater extent on some scales than others (e.g.,

Jean & Reynolds, 1982; Reynolds, 1998b), and is a more significant problem with personality scales than cognitive scales. Papers have even been published providing details on how to distort responses on personality tests in the desired direction (e.g. Whyte, 1967). Although there is no direct solution to this problem, many personality measures have built-in "lie" scales or social desirability scales to help detect deliberate faking to make one look as good as possible and F or infrequency scales to detect the faking of the presence of psychopathology.

The use and interpretation of scores from objective personality scales also has implications for this problem. Properly assessed and evaluated from an empirical basis, response to the personality scale is treated as the behavior of immediate interest and the actual content conveyed by the item becomes nearly irrelevant. As one example, there is an item on the Revised-Children's Manifest Anxiety Scale (RCMAS; Reynold's & Richmond, 1978, 1985), a test designed to measure chronic anxiety levels in children, that states "My hands feel sweaty." Whether the child's hands actually do feel sweaty is irrelevant. The salient question is whether children who respond "true" to this question are in reality more anxious than children who respond "false" to such a query. Children who respond more often in the keyed direction on the RCMAS display greater general anxiety and exhibit more observed behavior problems associated with anxiety than children who respond in the opposite manner. Although face validity of a personality or other test is a desirable quality, it is not always a necessary one. It is the actuarial implications of the behavioral response of choosing to respond in a certain manner that holds the greatest interest for the practitioner. Scales developed using such an approach are empirical scales.

Another approach is to devise content scales. As the name implies, the item content of such scales is considered more salient than its purely empirical relationship to the construct. Individuals with depression, especially men, may be edgy and irritable at times. Thus, the item "I sometimes lash out at others for no good reason" might show up on an empirically derived scale assessing depression, but is unlikely to find its way onto a content scale. "I am most often sad" would be a content-scale item assessing depression. Content scales are typically derived via expert judgments, but from an item pool that has passed muster at some empirical level already.

The emphasis on inner psychological constructs typical of personality scales poses special problems for their development and validation. A reasonable treatment of these issues can be

found in most basic psychological measurement texts (e.g., Anastasi & Urbina, 1997; Cronbach, 1983).

Objective personality scales are probably the most commonly used of all types of tests by clinical psychologists. They provide key evidence for the differentiation of various forms of psychopathology including clinical disorders and especially personality disorders. Omnibus scales such as the MMPI-2 and the Millon Clinical Multiaxial Inventory-3 are common with adult populations and also have adolescent versions. Omnibus scales directed at children and adolescents specifically however may be more appropriate for these younger age ranges. Among the many available, the Personality Inventory for Children and the Self-report of Personality from the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) are the most commonly used. Omnibus scales that are multidimensional in their construction are important to differential diagnosis in the early stages of contact with a patient. As diagnosis is established and treatment is in place, narrower scales that coincide with the clinical diagnosis become more useful for follow-up assessment and monitoring of treatment. In this regard, scales such as the Beck Depression Inventory or the State-Trait Anxiety Inventory are common examples. The tremendous cultural diversity of the world and how culture influences perceptions of items about the self and about one's family places special demands of cultural competence and cultural sensitivity on psychologists interpreting personality scales outside of their own cultural or ethnic heritage (e.g., see Dana, 1996; Hambleton, 1994; Moreland, 1995).

4.02.7.3 Projective Assessment

Projective assessment of personality has a long, rich, but very controversial history in the evaluation of clinical disorders and the description of normal personality. This controversy stems largely from the subjective nature of the tests used and the lack of good evidence of predictive validity coupled with sometimes fierce testimonial and anecdotal evidence of their utility in individual cases by devoted clinicians.

The subjectiveness of projective testing necessarily results in disagreement concerning the scoring and interpretation of responses to the test materials. For any given response by any given individual, competent professionals would each be likely to interpret differently the meaning and significance of the response. It is primarily the agreement on scoring that differentiates objective from subjective tests. If trained

examiners agree on how a particular answer is scored, tests are considered objective; if not, they are considered subjective. Projective is not synonymous with subjective in this context but most projective tests are closer to the subjective than objective end of the continuum of agreement on scoring of responses. Projective tests are sets of ambiguous stimuli, such as ink blots or incomplete sentences, and the individual responds with the first thought or series of thoughts that come to mind or tells a story about each stimulus. Typically no restrictions are placed on individuals' response options. They may choose to respond with anything desired; in contrast, on an objective scale, individuals must choose between a set of answers provided by the test or at least listed out for the examiner in a scoring manual. The major hypothesis underlying projective testing is taken from Freud (Exner, 1976). When responding to an ambiguous stimulus, individuals are influenced by their needs, interests, and psychological organization and tend to respond in ways that reveal, to the trained observer, their motivations and true emotions, with little interference from the conscious control of the ego. Various psychodynamic theories are applied to evaluating test responses, however, and herein too lie problems of subjectivity. Depending on the theoretical orientation of the psychologist administering the test, very different interpretations may be given. Despite the controversy surrounding these tests, they remain very popular.

Projective methods can be divided roughly into three categories according to the type of stimulus presented and the method of response called for by the examiner. The first category calls for the interpretation of ambiguous visual stimuli by the patient with an oral response. Tests in this category include such well-known techniques as the Rorschach and the Thematic Apperception Test (TAT). The second category includes completion methods, whereby the patient is asked to finish a sentence when given an ambiguous stem or to complete a story begun by the examiner. This includes the Despert Fables and a number of sentence completion tests. The third category includes projective art, primarily drawing techniques, although sculpture and related art forms have been used. In these tasks, the child is provided with materials to complete an artwork (or simple drawing) and given instructions for a topic, some more specific than others. Techniques such as the Kinetic-Family-Drawing, the Draw-A-Person, and the Bender-Gestalt Test fall in this category.

Criterion-related and predictive validity have proven especially tricky for advocates of projective testing. Although techniques such as the TAT are not as amenable to study and

validation through the application of traditional statistical and psychometric methods as objective tests may be, many clinical researchers have made such attempts with less than heartening results. None of the so-called objective scoring systems for projective devices has proved to be very valuable in predicting behavior, nor has the use of normative standards been especially fruitful. This should not be considered so surprising; however, it is indeed the nearly complete idiographic nature of projective techniques that can make them useful in the evaluation of a specific patient. It allows for any possible response to occur, without restriction, and can reveal many of a patient's current reasons for behaving in a specific manner. When used as part of a complete assessment, as defined in this chapter, projective techniques can be quite valuable. When applied rigidly and without proper knowledge and consideration of the patient's ecology, they can, as with other tests, be detrimental to our understanding of the patient. For a more extensive review of the debates over projective testing, the reader is referred to Exner (1976), Jackson and Messick (1967, Part 6), O'Leary and Johnson (1979), and Prevatt (in press), as well as to other chapters in this volume, especially Chapter 15.

4.02.7.4 Behavioral Assessment

The rapid growth of behavior therapy and applied behavior analysis has led to the need for tests that are consistent with the theoretical and practical requirements of these approaches to the modification of human behavior. Thus, the field of behavioral assessment has developed and grown at an intense pace. Book length treatments of the topic became commonplace in the 1970s and 1980s (e.g., Haynes & Wilson, 1979; Hersen & Bellack, 1976; Mash & Terdal, 1981) and entire journals are now devoted to research regarding behavioral assessment (e.g., *Behavioral Assessment*). The general term "behavioral assessment" has come to be used to describe a broad set of methods including some traditional objective personality scales, certain methods of interviewing, physiological response measures, naturalistic observation, norm-referenced behavior checklists, frequency counts of behavior, and a host of informal techniques requiring the observation of a behavior with recording of specific responses. Behavioral assessment will be discussed here in its more restricted sense to include the rating (by self or others) of observable behavioral events, primarily taking the form of behavior checklists and rating forms that may or may not be normed. Although I would certainly include psychophysiological assessment within this

category, the scope of the work simply will not allow us to address this aspect of behavioral assessment except to say that it is indeed a most useful one in the treatment of a variety of clinical disorders.

The impetus for behavioral assessment comes not only from the field of behavior therapy but also from a general revolt against the high level of inference involved in such methods of assessing behavior as the Rorschach and the TAT. The greatest distinguishing characteristic between the behavioral assessment of psychopathological disorders and most other techniques is the level of inference involved in moving from the data provided by the assessment instrument to an accurate description of the patient and the development of an appropriate intervention strategy. This is a most useful strength for behavioral assessment strategies but is their greatest weakness when it is necessary to understand what underlies the observed behaviors.

Many of the early conceptual and methodological issues have been resolved in this area of assessment, for example, the importance of norms and other traditional psychometric concepts such as reliability and validity (Cone, 1977; Nelson, Hay, & Hay, 1977). Problems of interobserver reliability and observer drift remain but are well on their way to being resolved. Unquestionably, behavioral assessment is an exciting and valuable part of the assessment process. Behavioral assessment grew from a need to quantify observations of a patient's current behavior and its immediate antecedents and consequences, and this is the context within it that remains most useful today.

There are a number of formal behavior rating scales or behavior checklists now available. These instruments typically list behaviors of interest in clearly specified terms and have a trained observer or an informant indicate the frequency of occurrence of these behaviors. Interpretation can then take on a normative or a criterion-reference nature depending on the purpose of the assessment and the availability of norms. Clusters of behaviors may be of interest that define certain clinical syndromes such as attention deficit hyperactivity disorder. On the other hand, individual behaviors may be the focus (e.g., hitting other children). More frequently, behavioral assessment occurs as an "informal" method of collecting data on specific behaviors being exhibited by a patient and is dictated by the existing situation into which the psychologist is invited. An informal nature is dictated by the nature of behavioral assessment in many instances. Part of the low level of inference in behavioral assessment lies in not generalizing observations of behavior across

settings without first collecting data in multiple settings. In this regard, behavioral assessment may for the most part be said to be psychosituational. Behavior is observed and evaluated under existing circumstances, and no attempt is made to infer that the observed behaviors will occur under other circumstances. Comprehensive systems that are multimethod, multidimensional, and that assess behavior in more than one setting have been devised and provide a good model for the future (Reynolds & Kamphaus, 1992).

Another area of assessment that stems from behavioral psychology and is considered by many to be a subset of behavioral assessment is task analysis. Whereas behavioral assessment typically finds its greatest applicability in dealing with emotional and behavioral difficulties, task analysis is most useful in evaluating and correcting learning problems. In task analysis, the task to be learned (e.g., brushing one's teeth or multiplying two-digit numbers) is broken down into its most basic component parts. The learner's skill at each component is observed and those skills specifically lacking are targeted for teaching to the child. In some cases, hierarchies of subskills can be developed, but these have not held up well under cross-validation. Task analysis can thus be a powerful tool in specifying a learner's existing (and needed) skills for a given learning problem. Task analysis could, for example, form an integral part of any behavioral intervention for a child with specific learning problems. The proper use of these procedures requires a creative and well-trained individual conversant with both assessment technology and behavioral theories of learning, since there are no standardized task analysis procedures. Those involved in task analysis need to be sensitive to the reliability and validity of their methods. As with other behavioral assessment techniques, some contend that behavioral assessment techniques need only demonstrate that multiple observers can agree on a description of the behavior and when it has been observed. Though not having to demonstrate a relationship with a hypothetical construct, behavioral techniques must demonstrate that the behavior observed is consistent and relevant to the learning problems. For behavior checklists and more formal behavioral assessment techniques, most traditional psychometric concepts apply and must be evaluated with regard to the behavioral scale in question.

4.02.7.5 Neuropsychological Assessment

Along with behavioral assessment, perhaps the most rapidly growing area in patient evaluation is neuropsychological assessment. Many view neuropsychological assessment as

the application of a specific set of tests or battery of tests. Far from being a set of techniques, the major contribution of neuropsychology to the assessment process is the provision of a strong paradigm from which to view assessment data (Reynolds, 1981b, 1981c, 1997). Without a strong theoretical guide to test score interpretation, one quickly comes to rely upon past experience and illusory relationships and trial and error procedures when encountering a patient with unique test performance. As with most areas of psychology, there are competing neuropsychological models of cognitive functioning, any one of which may be most appropriate for a given patient. Thus considerable knowledge of neuropsychological theory is required to evaluate properly the results of neuropsychological testing.

Since the 1950s, clinical testing in neuropsychology has been dominated by the Halstead-Reitan Neuropsychological Test Battery (HRNTB), although the Luria-Nebraska Neuropsychological Battery and the Boston process approach have made significant inroads. The prevalence of use of the HRNTB is partly responsible for perceptions of clinical neuropsychology as primarily a set of testing techniques. However, a brief examination of the HRNTB should quickly dispel such ideas. The HRNTB consists of a large battery of tests taking a full day to administer. There is little that can be said to be psychologically or psychometrically unique about any of these tests. They are all more or less similar to tests that psychologists have been using for the past 50 years. The HRNTB also typically includes a traditional intelligence test such as one of the Wechsler scales. The HRNTB is unique in the particular collection of tests involved and the method of evaluating and interpreting performance. While supported by actuarial studies, HRNTB performance is evaluated by the clinician in light of existing neuropsychological theories of cognitive function, giving the battery considerable explanatory and predictive power.

Neuropsychological approaches to clinical assessment are rapidly growing and can be most helpful in defining areas of cognitive-neuropsychological integrity and not just in evaluating deficits in neurological function. Neuropsychological techniques can also make an important contribution by ruling out specific neurological problems and pointing toward environmental determinants of behavior. The well-trained neuropsychologist is aware that the brain does not operate in a vacuum but is an integral part of the ecosystem of the patient. As with other methods of assessment, neuropsychological assessment has much to offer the assessment process when used wisely; poorly or carelessly

implemented, it can create seriously false impressions, lessen expectations, and precipitate a disastrous state of affairs for the patient it is designed to serve. Clinicians who use neuropsychological approaches to their work or make neuropsychological interpretations of tests or behaviors are in high demand but require specialized training that takes years to acquire.

4.02.8 CLINICAL VS. STATISTICAL PREDICTION

Given a set of test results and/or additional interview or historical data on a patient, there are two fundamental approaches a clinician can apply to diagnostic decision-making or to the prediction of future behavior. The first, and likely most common, is the simple human judgment of the clinician who reviews the data in conjunction with prior experiences and knowledge of diagnostic criteria, psychopathology, and other psychological literature. As a result of the application of this *clinical method*, a diagnosis or prediction is made. Alternatively, the clinician may apply a formal algorithm or set of mathematical equations to predict membership in a class (a diagnosis) or the probability of some future behavior (e.g., whether a defendant will or will not molest children again if placed on probation). The use of such mechanistic approaches constitutes the *statistical method*.

The ability of clinicians to use test data in optimal ways and to combine multiple sources of data optimally to arrive at correct diagnoses and to make predictions about future behavior has been the subject of much scrutiny. Meehl (1954) addressed this problem early on and concluded that formula-based techniques, derived by mathematical models that search for optimal combinations of data, are superior to clinicians in decision-making. This has been difficult for clinicians to accept even as more and more actuarial systems for test interpretation find their way into our office computers. In 70+ years of research on this topic, actuarial modeling continues to be superior (Faust & Ackley, 1998; Grove & Meehl, 1996), yet I listened to a clinical psychologist testify in February of 1998 that clinical judgment was always better and that actuarial predictions were used only when you had nothing else to use. In 136 studies since 1928, over a wide range of predictions, actuarial modeling is invariably equal to or superior to clinical methods (Grove & Meehl, 1996). Grove and Meehl (1996) have addressed this reluctance (or perhaps ignorance) about actuarial modeling in clinical-

decision-making and clinicians' objections to acceptance of more than 70 years of consistent research findings, as has Kleinmuntz (1990), who did seminal research on developing expert systems for MMPI interpretation in the 1960s.

Grove and Meehl (1996) state the most common objection of clinicians to statistical modeling is that they (the clinicians) use a combination of clinical and statistical methods, obviating the issue since they are used in a complementary model. This is a spurious argument because, as research shows, the clinical method and the statistical method often disagree and there is no known way to combine the two methods to improve predictions; we simply do not know under what conditions to conclude the statistical model will be wrong (also see Faust & Ackley, 1998, and Reynolds, 1998a). Grove and Meehl (1996) illustrate this quandary by examining the actions of an admissions committee. Suppose the applicant's test scores, grade point average, and rank in class predict successful college performance but the admissions commission believes, perhaps based on an interview and letters of recommendation, that the applicant will not be successful. The two methods cannot be combined when they specify different outcomes. One is right and one is wrong. Consider a crucial prediction in a forensic case. A psychologist treating an offender on parole for aggravated sexual assault of a child is asked whether the offender might molest a child of the same age and gender as a prior victim if the child is placed in the offender's home (he recently married the child's mother). Actuarial tables with rearrest rates for offenders with many similar salient characteristics and in treatment indicate that 10–11% of those individuals will be arrested again for the same offense. The psychologist concludes that the offender is virtually a zero percent risk, however, because "he has worked so hard in treatment." The two methods again cannot be resolved; yet, as clinicians we persist in believing we can do better than the statistical models even in the face of a substantial body of contradictory evidence. Grove and Meehl (1996) review some 16 other objections clinicians make to statistical methods in diagnosis and other clinical predictions. Most resolve to questions of validity, reliability, or cost. Cost and inconveniences are rapidly becoming a nonissue as computerized models are widely available and some of these cost less than one dollar per application for a patient (e.g., Stanton, Reynolds, & Kamphaus, 1993).

Statistical models work better, consistently, and for known reasons. Clinicians cannot assign optimal weights to variables they use to make decisions, do not apply their own rules

consistently, and are influenced by relatively unreliable data (e.g., see Dawes, 1988; Faust, 1984; Grove & Meehl, 1996; Meehl, 1954). As the reliability of a predictor goes down, the relative accuracy of any prediction will be reduced and, consequently, the probability of being wrong increases. The decisions being made are not trivial. Every day, thousands of diagnostic decisions are being made that influence treatment choices, freedom for parolees and probationers, large monetary awards in personal injury cases, custody of children, and others. What clinical psychologists do is important and the increased use of statistical models based on sound principles of data collection that includes data from well-standardized, objective psychological tests seems imperative from an ethical standpoint and from the standpoint of the survival of the profession as accountability models are increasingly applied to reviews of the need for and effectiveness of our services.

4.02.9 ACCESSING CRITICAL COMMENTARY ON STANDARDIZED PSYCHOLOGICAL TESTS

Not every practitioner can nor should be an expert on the technical adequacy and measurement science underlying each and every standardized test that might be useful to their clinical practice. With a clear understanding of the fundamentals of measurement, however, clinicians can make intelligent choices about test selection and test use based on the test manuals and accompanying materials in most cases. However, when additional expertise or commentary is required, critical reviews of nearly every published test can be accessed with relative ease.

Many journals in psychology routinely publish reviews of new or newly revised tests including such publications as the *Archives of Clinical Neuropsychology*, *Journal of Psychoeducational Assessment*, *Journal of Learning Disabilities*, *Journal of School Psychology*, *Child Assessment News*, and the *Journal of Personality Assessment*. However, the premier source of critical information on published tests are the publications of the Buros Institute of Mental Measurement.

In the late 1920s, Oscar Krisen Buros began to publish a series of monographs reviewing statistics texts. He noted the rapid proliferation of psychological and educational tests beginning to occur at the same time and rapidly turned his attention to obtaining high-quality reviews of these tests and publishing them in

bound volumes known as the *Mental Measurements Yearbook* (MMY). The first MMY was published by Buros in 1938 and the series continues today. Buros died in 1978, during the final stages of production of the Eighth MMY (though “Yearbooks,” they are not published annually), and his spouse, art director, and assistant Luella Buros saw the Eighth MMY to completion. Subsequently, she opened a competition for proposals to adopt the Institute and continue the work of her late husband. A proposal written by this author (then on the faculty of the University of Nebraska-Lincoln) was chosen and the Buros Institute of Mental Measurement was established in 1979 at the University of Nebraska-Lincoln, where it remains permanently due to a generous endowment from Luella Buros.

The Institute continues to seek out competent reviewers to evaluate and provide critical commentary on all educational and psychological tests published in the English language. These reviews are collected in an ongoing process, as tests are published or revised, under a strict set of rules designed to ensure fair reviews and avoid conflicts of interest. The collected reviews are published on an unscheduled basis approximately every five to eight years. However, as reviews are written and accepted for publication, they are added quickly to the Buros Institute database which may be searched on-line by subscription to the master database or through most major university libraries. Information on how to access current reviews in the Buros database can be obtained through nearly any reference librarian or through a visit to the Buros Institute website. The Institute established a sterling reputation as the “consumer reports” of the psychological testing industry under the 50 year leadership of Oscar Buros and this reputation and service have been continued at the University of Nebraska-Lincoln. Consumers of tests are encouraged to read the Buros reviews of the tests they choose to use.

4.02.10 CONCLUDING REMARKS

Knowledge of measurement science and psychometrics is critical to good practice in clinical assessment. Hopefully, this review, targeted at practice, has provided a better foundation for reading about psychological testing and for developing better skills in the application of methods of testing and assessment. Old tests continue to be revised and updated, and many new tests are published yearly. There is a large, rapidly growing body of literature on test interpretation that is too often

ignored in practice (e.g., Reynolds & Whitaker, in press) but must be accessed in practice. It is difficult but necessary to do so. Measurement science also progresses and practitioners are encouraged to revisit their mathematical backgrounds every few years as part of what has become a nearly universal requirement for continuing education to continue in practice. New paradigms will always emerge as they have in the past.

It is from basic psychological research into the nature of human information processing and behavior that advances in psychological assessment must come. While some of these advances will be technological, the more fruitful area for movement is in the development of new paradigms of test interpretations. With each advance, with each "new" test that appears, we must proceed with caution and guard against jumping on an insufficiently researched bandwagon. Fruitful techniques may be lost if implemented too soon to be fully understood and appreciated; patients may also be harmed by the careless or impulsive use of assessment materials that are poorly designed (but attractively packaged) or without the necessary theoretical and empirical grounding. When evaluating new psychological assessment methods, surely *caveat emptor* must serve as the guard over our enthusiasm and our eagerness to provide helpful information about patients in the design of successful intervention programs.

4.02.11 REFERENCES

- American Psychological Association (1985). *Standards for educational and psychological tests*. Washington, DC: Author.
- Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.
- Anastasi, A. (1981). *Psychological testing* (5th ed.). New York: Macmillan.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.
- Barrios, B. A., Hartmann, D. P., & Shigetomi, C. (1981). Fears and anxieties in children. In E. J. Marsh & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders*. New York: Guilford.
- Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 8, 411–426.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.
- Cronbach, L. J. (1983). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137–163.
- Dana, R. H. (1996). Culturally competent assessment practices in the United States. *Journal of Personality Assessment*, 66, 472–487.
- Dawes, R. M. (1988). *Rational choice in an uncertain world*. Chicago, IL: Harcourt Brace Jovanovich.
- Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Exner, J. E. (1976). Projective techniques. In I. B. Weiner (Ed.), *Clinical methods in psychology*. New York: Wiley.
- Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis, MN: University of Minnesota Press.
- Faust, D., & Ackley, M. A. (1998). Did you think it was going to be easy? Some methodological suggestions for the investigation and development of malingering detection techniques. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 1–54). New York: Plenum.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 105–146). New York: Macmillan.
- Grove, W. M., & Meehl, P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, 2, 293–323.
- Hambleton, R. K. (1994). Guidance for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–244.
- Haynes, S. N., & Wilson, C. C. (1979). *Behavioral assessment*. San Francisco: Jossey-Bass.
- Hays, W. L. (1973) *Statistics for the social sciences*. New York: Holt, Rinehart & Winston.
- Hersen, M., & Bellack, A. S. (1976). *Behavioral assessment: A practical handbook*. New York: Pergamon.
- Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York: Plenum.
- Jackson, D. N., & Messick, S. (Eds.) (1967). *Problems in human assessment*. New York: McGraw-Hill.
- Jean, P. J., & Reynolds, C. R. (1982). *Sex and attitude distortions: The faking of liberal and traditional attitudes about changing sex roles*. Paper presented to the annual meeting of the American Educational Research Association, New York, March.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Kaufman, A. S. (1990). *Assessment of adolescent and adult intelligence*. Boston: Allyn & Bacon.
- Kaufman, A. S. (1994) *Intelligent testing with the WISC-III*. New York: Wiley.
- Kaufman, A. S., & Reynolds, C. R. (1983). Clinical evaluation of intellectual function. In I. Weiner (Ed.), *Clinical methods in psychology* (2nd ed.). New York: Wiley.
- Kleimuntz, B. (1990). Why we still use our heads instead of the formulas: Toward an integrative approach. *Psychological Bulletin*, 107, 296–310.
- Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mash, E. J., & Terdal, L. G. (1981). *Behavioral assessment of childhood disorders*. New York: Guilford.
- Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012–1027.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.
- Moreland, K. L. (1995). Persistent issues in multicultural assessment of social and emotional functioning. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (Eds.), *Handbook of multicultural Assessment: Clinical, psycho-*

- logical, and educational applications. San Francisco: Jossey-Bass.
- Nelson, R. O., Hay, L. R., & Hay, W. M. (1977). Comments on Cone's "The relevance of reliability and validity for behavioral assessment." *Behavior Therapy*, 8, 427-430.
- Newland, T. E. (1980). Psychological assessment of exceptional children and youth. In W. M. Cruickshank (Ed.), *Psychology of exceptional children and youth* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
- O'Leary, K. D., & Johnson, S. B. (1979). Psychological assessment. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood*. New York: Wiley.
- Palmer, D. J. (1980). Factors to be considered in placing handicapped children in regular classes. *Journal of School Psychology*, 18, 163-171.
- Petersen, N. S., Kolen, M., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. Linn (Ed.) *Educational measurement* (3rd ed., pp. 221-262). New York: Macmillan.
- Prevatt, F. (in press). Personality assessment in the schools. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed.). New York: Wiley.
- Reynolds, C. R. (1981a). The fallacy of "two years below grade level for age" as a diagnostic criterion for reading disorders. *Journal of School Psychology*, 19, 350-358.
- Reynolds, C. R. (1981b). Neuropsychological assessment and the habilitation of learning: Considerations in the search for the aptitude X treatment interaction. *School Psychology Review*, 10, 343-349.
- Reynolds, C. R. (1981c). The neuropsychological basis of intelligence. In G. Hynd & J. Obrzut (Eds.), *Neuropsychological assessment of the school-aged child*. New York: Grune and Stratton.
- Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology*. New York: Wiley.
- Reynolds, C. R. (1985). Critical measurement issues in learning disabilities. *Journal of Special Education*, 18, 451-476.
- Reynolds, C. R. (1997). Measurement and statistical problems in neuropsychological assessment of children. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *The handbook of clinical child neuropsychology* (2nd ed., pp. 182-203). New York: Plenum.
- Reynolds, C. R. (1998a). Common sense, clinicians, and actuarialism in the detection of malingering during head injury litigation. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 261-282). New York: Plenum.
- Reynolds, C. R. (Ed.) (1998b). *Detection of malingering during head injury litigation*. New York: Plenum.
- Reynolds, C. R. (in press-a). Need we measure anxiety separately for males and females? *Journal of Personality Assessment*.
- Reynolds, C. R. (in press-b). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy and Law*.
- Reynolds, C. R. & Brown, R. T. (1984). Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing*. New York: Plenum.
- Reynolds, C. R., & Clark, J. (1982). Cognitive assessment of the preschool child. In K. D. Paget & B. Bracken (Eds.), *Psychoeducational assessment of the preschool and primary aged child*. New York: Grune & Stratton.
- Reynolds, C. R., & Kamphaus, R. W. (Eds.) (1990a). *Handbook of psychological and educational assessment of children: Vol I. Intelligence and achievement*. New York: Guilford.
- Reynolds, C. R., & Kamphaus, R. W. (Eds.) (1990b). *Handbook of psychological and educational assessment of children: Vol II. Personality, behavior, and context*. New York: Guilford.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Service.
- Reynolds, C. R., & Richmond, B. O. (1978). What I think and feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology*, 6, 271-280.
- Reynolds, C. R., & Richmond, B. O. (1985). *Revised children's manifest anxiety scale*. Los Angeles: Western Psychological Services.
- Reynolds, C. R., & Whitaker, J. S. (in press). Bias in mental testing since Jensen's "Bias in mental testing." *School Psychology Quarterly*.
- Salvia, J., & Ysseldyke, J. E. (1981). *Assessment in special and remedial education* (2nd ed.). Boston: Houghton Mifflin.
- Sattler, J. (1988). *Assessment of children's intelligence and special aptitudes*. San Diego, CA: Author.
- Stanton, H., Reynolds, C. R., & Kamphaus, R. W. (1993). *BASC plus software scoring program for the Behavior Assessment System for Children*. Circle Pines, MN: American Guidance Service.
- Weschler, D. (1992). *Weschler Intelligence Scale for Children-III*. San Antonio, TX: Psychological Corporation.
- Whyte, W. H. (1967). How to cheat on personality tests. In D. Jackson & S. Messick (Eds.), *Problems in human assessment*. New York: McGraw-Hill.

4.03

Diagnostic Models and Systems

ROGER K. BLASHFIELD

Auburn University, AL, USA

4.03.1 INTRODUCTION	57
4.03.2 PURPOSES OF CLASSIFICATION	58
4.03.3 DEVELOPMENT OF CLASSIFICATION SYSTEMS IN THE USA	59
4.03.4 KRAEPLIN	62
4.03.5 EARLY <i>DSMS</i> AND <i>ICDS</i>	62
4.03.6 NEO-KRAEPELINIANS	63
4.03.7 <i>DSM-III</i> , <i>DSM-III-R</i> , AND <i>DSM-IV</i>	64
4.03.8 <i>ICD-9</i> , <i>ICD-9-CM</i> , AND <i>ICD-10</i>	67
4.03.9 CONTROVERSIES AND ISSUES	69
4.03.9.1 <i>Organizational Models of Classification</i>	69
4.03.9.2 <i>Concept of Disease</i>	72
4.03.9.3 <i>Two Views of a Hierarchical System of Diagnostic Concepts</i>	75
4.03.9.4 <i>Problem of Diagnostic Overlap/Comorbidity</i>	76
4.03.10 CONCLUDING COMMENTS	78
4.03.11 REFERENCES	78

4.03.1 INTRODUCTION

To all the People and Inhabitants of the United States and all the outlying Countries, Greetings: I, John Michler, King of Tuskaroras, and of all the Islands of the Sea, and of the Mountains and Valleys and Deserts; Emperor of the Diamond Caverns, and Lord High General of the Armies thereof; First Archduke of the Beautiful Isles of the Emerald Sea; Lord High Priest of the Grand Lama, etc., etc., etc.: Do issue this my proclamation. Stand by and hear, for the Lord High Shepherd speaks. No sheep have I to lead me around, no man have I to till me the ground, but the sweet little cottage is all of my store, and my neat little cottage has ground for the floor. No children have I to play me around, no dog have I to bark me around, but the three-legged stool is the chief of my store, and my neat little cottage has ground for the floor.

Yea, verily, I am the Mighty King, Lord Archduke, Pope, and Grand Sanhedrim, John Michler. None with me compare, none fit to comb my hair, but the three-legged stool is the chief of my store, and my neat little cottage has ground for the floor. John Michler is my name. Selah!

I am the Great Hell-Bending Rip-Roaring Chief of the Aborigines! Hear me and obey! My breath overthrows mountains; my mighty arms crush the everlasting forests into kindling wood; I am the owner of the Ebony Plantations; I am the owner of all the mahogany groves and of all the satin-wood; I am the owner of all the granite; I am the owner of all the marble; I am the owner of all the owners of Everything. Hear me and obey! I, John Michler, stand forth in the presence of the Sun and of all the Lord Suns and Lord Planets of the Universe, and I say, Hear me and obey! I, John Michler, on this eighteenth day of August, 1881, do say, Hear me and obey! for with me none can equal, no, not one, for the three-legged stool is the chief of my store,

and my neat little cottage has ground for the floor. Hear me and obey! Hear me and obey! John Michler is my name.

John Michler, First Consul and Dictator of the World, Emperor, Pope, King and Lord High Admiral, Grand Liconthropon forever! (Hammond, 1883, pp. 603–604)

A physician in private practice in New York City reported that a man brought his brother, John Michler, to see him. John was acting strangely, and his brother wanted to know what to do. The brother gave the physician a proclamation that John Michler had written (see above). Clearly, to most observers, there would be no question that John Michler was “crazy.” However, what is the diagnosis of John Michler?

When this proclamation was shown to mental health professionals, the most common diagnostic possibilities that are mentioned are schizophrenia and bipolar disorder (manic episode). What do these various diagnoses mean? Why did clinicians not assign a diagnosis of narcissistic personality disorder to this patient? Certainly this man would fit the vernacular meaning of self-centered and self-aggrandizing that is often associated with a narcissistic personality. How is a manic episode differentiated from schizophrenia? What does it mean to say that Michler appears to be psychotic? Does that diagnosis mean that he has some type of disease that has affected part of his brain, or does it suggest that his childhood was so unusual and abnormal that he has developed a strange and unusual way of coping with life?

4.03.2 PURPOSES OF CLASSIFICATION

The vernacular word that was used to describe John Michler was “crazy.” This word is frequently used in descriptions of persons who have mental disorders. The reason for the applicability of this word is that one common feature of most psychiatric patients is that their behaviors are statistically abnormal. That is, psychiatric patients behave in ways that are deviant. Their interpersonal actions are not expected within the cultural context of our society.

Classification is a fundamental human activity that people use to understand their world. For instance, a classification of animals is helpful when understanding the variations among diverse forms of living organisms. In forming a classification, the classifier is attempting to use observed similarities and differences among the things being classified, so as to find some order among those things. In psychopathology, the general goal of classification is an attempt to use similarities and differences among people who

behave in deviant and socially abnormal ways in order to understand their behaviors.

More specifically, there are five purposes to the classification of mental disorders: (i) forming a nomenclature so that mental health professionals have a common language; (ii) serving as a basis of information retrieval; (iii) providing a short-hand description of the clinical picture of the patient; (iv) stimulating useful predictions about what treatment approach will be best; and (v) serving as a concept formation system for a theory (or theories) of psychopathology.

The first reason to have a classification system, providing a nomenclature, is the most fundamental (World Health Organization, 1957). At a minimum, a classification system provides a set of nouns for clinicians to use to discuss their objects of interest—people. Thus, a nomenclature is a set of terms that refer to groups of people that mental health professionals see in their various professional roles.

The second reason, information retrieval, has a pragmatic focus on how well a classification organizes a research literature, so that clinicians and scientists can search for information that they need (Feinstein, 1967). In biology, there is a dictum: “the name of a plant is the key to its literature” (Sneath & Sokal, 1973). The same is true in the area of mental disorders. If a student clinician is assigned a patient who attempts to control weight by inducing vomiting, the name bulimia becomes important for helping the student locate the literature about this disorder in books, journal articles, and even on the internet.

The third reason for having a classification is description. There are many ways of creating classifications that could satisfy the first two purposes. For instance, clinicians could decide to classify all of their patients on eye color. Using eye color would allow clinicians to have a language to discuss patients (“I have seen 17 brown eyed, eight blue eyed, and four mixed eye color patients in the last month.”). Also, eye color categories could be used as names to store information about patients. However, using eye color to classify patients would not be a satisfactory solution to either researchers or clinicians because patients with similar eye colors are unlikely to have similar symptom patterns. To meet the purpose of description, patients are grouped on the basis of similarity. Patients who have the same diagnosis are expected to be relatively similar in terms of symptoms (Lorr, 1966). In addition, these patients should be dissimilar when compared to patients with different diagnoses (Everitt, 1974). In the case of John Michler, if he is having a manic episode of a bipolar disorder, then we would expect that Michler’s brother would

report that John had been spending large sums of money that he did not have, that his speech was extremely rapid, and that his sleep pattern was markedly disturbed. In contrast, his brother is unlikely to report that Michler usually sat around the house with an emotionless, cold, detached interpersonal style and that he would tell his brother about voices in his head that were in discussion about Michler's behaviors. The latter symptoms typically occur in individuals with schizophrenia. Thus, diagnoses become descriptive short-hand names for clusters of co-occurring symptoms.

The fourth purpose is prediction. This purpose typically involves two types of inferences: (i) predicting the course of the patient's condition if there is no treatment or intervention ("Diagnosis is prognosis" as stated by Woodruff, Goodwin, & Guze, 1974); and (ii) predicting which treatment approach will be most effective with the patient (Roth & Fonagy, 1996). In the field of psychopathology, prediction has proved to be an elusive goal. Recently, there was an important multisite study that was performed in the USA in which three different treatment approaches to alcoholism were compared. An attempt was made to see whether particular groups of patients improved with specific treatments. The initial results have been negative. Except for differences related to the severity of patient symptomatology, other patient characteristics did not predict which treatments were most effective (Project Match Research Group, 1997).

The final goal of a classification is concept formation (Hempel, 1965). This goal is probably the most distant. In biological classification, Linneaus and his contemporaries made notable gains in the classification of living organisms by creating classifications that served to describe most of the known information about these organisms. Almost a century later, Darwin formulated a theory of evolution which explained many of the organizational phenomena that Linneaus had used (Hull, 1988). In particular, Darwin's evolutionary theory provided a basis for understanding why there is a hierarchical arrangement about the categories in biological classification. The field of biological classification has continued to change. During the twentieth century, three different, competing approaches to biological classification have appeared (Hull, 1988). In oversimplified terms, one of these approaches emphasized the nomenclature and information retrieval purposes, the second focused on description, and the third was based on a theoretical view. The third view, the one based on theory, has become the dominant approach to biological classification (Nelson & Platnick, 1981).

4.03.3 DEVELOPMENT OF CLASSIFICATION SYSTEMS IN THE USA

The classification of mental disorders has an extensive history that can be traced back to some of the earliest writings known to man. A nineteenth century BC Egyptian writer discussed a disorder in women in which they would report vague and inconsistent physical symptoms that appeared to shift in body location over time (Veith, 1965). Psalm 102 provides an excellent clinical description of depression.

However, like many others areas of modern science, the first major commentaries on mental disorders were found in the writings of the Greeks. The Greek medical writers introduced four terms, all of which are still used today: "melancholia," "hysteria," "mania," and "paranoia." Melancholia was a Greek term that referred to a condition that now would be described by the word depression. Hippocrates believed that the sadness and the slowed bodily movements associated with this disorder were caused by an abundance of black bile, which he considered to be one of the four main ingredients in the human body. Hence, he named this disorder *melan* (black) + *cholia* (bile). The second term, hysteria, was the Greek word for the condition originally described by the Egyptians in which women had multiple, inconsistent and changing somatic complaints. The Hippocratic writers used the name of hysteria, which means pertaining to the uterus, because they believed that this disorder was caused by a dislodged, floating uterus. The last two terms were mania and paranoia. Mania, to the Greeks, referred to persons who were delusional. During the nineteenth century, individuals who were delusional but had few other symptoms were diagnosed with monomania (Berrios, 1993). Mania became an umbrella term for almost any type of psychotic state (Spitzka, 1883). The meaning of mania, however, changed again in the twentieth century to its contemporary denotation of grandiosity, excitement, expansiveness, and elation. The final Greek term, paranoia, has undergone a similar transformation. Paranoia initially meant craziness (*para* = abnormal + *nous* = mind). Now the term refers to people who are suspicious and often have delusions in which others are plotting against them.

After the Greeks, psychopathology did not attract much scientific interest until the nineteenth century. During the Middle Ages, mental disorders were associated with evilness. Thus, mental disorders were under the domain of religious authorities, not physicians or scientists. This approach to mental disorders began

to change in the late 1700s as exemplified by the fact that King George III of England, who was psychotic for most of the last decade of his reign, received medical care rather than religious counseling.

The first major American physician to be interested in mental disorders was Benjamin Rush who, as a signer of the Declaration of Independence, was one of the prominent American physicians of the late eighteenth century. He was very interested in the forms of insanity. Rush also published a book on the topic which he titled *Medical inquires and observations upon the diseases of the mind* (Alexander & Selesnick, 1966). He believed in a theory of neurosis. According to this theory, mental disorders were caused by overstimulation of the nervous system. Thus, environmental phenomena such as the pace of urban living, overuse of alcohol, excessive sexual behavior, masturbation, and smoking were all seen as casual factors in the development of mental disorders. As a result, asylums were the appropriate way to treat insane patients. Asylums could provide the quiet and tranquility that was necessary to allow the nervous system to heal and to repair itself.

About the same time that Rush was writing about psychopathology in the USA, there was an important discovery in France that was to markedly influence thinking about mental disorders. In 1822, a physician named Bayle (Quetel, 1990) performed autopsies on a number of patients who presented with gradiose delusions and dementia (i.e., who had lost their mind from the French *de-* (not) + *ment* (mind)). Bayle discovered that all of the patients in his study had marked changes in their brains. In addition to their dementia, all of these patients developed motor paralysis before they died. The brains of these patients had shrunk to almost half the weight of a normal brain, the skin of the brain (i.e., the meninges and the arachnoid) was thickened, and the color of the brain was strikingly different from that of normal brains. Bayle's name for this disorder was chronic arachnitis since he believed that this disorder was caused by a chronic infection of arachnoid tissue (Quetel, 1990). Later, the common name for this disorder was changed to dementia paralytica, a descriptive name that referred to the combined presence of a psychosis together with the progressive paralysis of the patient's limbs.

The discovery of dementia paralytica was the first instance in which a mental disorder had been shown to be associated with demonstrable changes in the brain. A number of autopsy studies appeared in the medical journals. These studies attempted to identify the exact neuro-

pathology of this disorder as well as to understand its cause. Dementia paralytica was also a clinically serious disorder because it accounted for about one-sixth of all admissions to insane asylums during the nineteenth century. The prognosis of the disorder was very poor because death typically would occur within three years of the diagnosis (Austin, 1859/1976). For most of the nineteenth century, many different etiologies were proposed as potential causes of this disorder. Austin (1859/1976), for instance, listed the following as moral causes of dementia paralytica: death of a son, sudden loss of two good situations, wife's insanity, worry, and commercial ruin.

With the increasing interest in psychopathology during the nineteenth century, a number of classifications for mental disorders appeared. One example of these classifications was published by William A. Hammond (Blustein, 1991). Hammond, like Freud, was a nineteenth century neurologist. As a young physician, Hammond had published a set of interesting experimental studies on human physiology. At the age of 34, he became Surgeon General for the US Army during the Civil War and was credited with a number of important innovations at the time including hospital design, the development of an ambulance corps, and the removal of a mercury compound from the medical formulary. His political clashes with the Secretary of War, however, led to his court martial and dismissal. After the Civil War, he moved to New York City and developed a lucrative private practice as a neurologist—a remarkable accomplishment during a time when most physicians were generalists. His interests extended to psychiatry and to writing novels as well as to physiology, studies of sleep, hypnosis, and the use of animal hormonal extracts. Hammond wrote extensively in scientific journals. He was one of the founders of the *Journal of Nervous and Mental Disease* which is still published. In addition, he authored important textbooks of neurology and psychiatry.

In Hammond's textbook of mental disorders, he argued that there were six possible principles that could be used to organize a classification system:

- (i) anatomical (organized by the part of the brain that is affected);
- (ii) physiological (organized by the physiological system in the brain);
- (iii) etiological (supposed causes);
- (iv) psychological (based upon a functional view of the mind);
- (v) pathological (observable, morbid alterations in the brain); and
- (vi) clinical (descriptive, based upon clusters of symptoms).

Of these six principles, Hammond said that the anatomical, the physiological, and the pathological are the best, but he could not use them because the science of his time was insufficient. Hammond also rejected the etiological organization of categories, because he felt that an etiological classification, given nineteenth century knowledge, would be too speculative. Thus, the main choice was between the clinical (descriptive) approach and the psychological (mentalistic) approach. Hammond preferred the latter because he thought that a classification which did not have a strong theoretical basis would fail.

Hammond adopted a functional view of psychology that was common in his day. He believed that mental functioning could be organized into four areas: perception, cognition, affect, and will. Hence, he organized his classification of mental disorders into six major headings:

- (i) perceptual insanities (e.g., hallucinations);
- (ii) intellectual insanities (e.g., delusional disorders);
- (iii) emotional insanities (e.g., melancholia);
- (iv) volitional insanities (e.g., abulomania);
- (v) compound insanities (i.e., disorders affecting more than one area of the mind); and
- (vi) constitutional insanities (i.e., disorders with specific causes such as choreic insanity).

There were a total of 47 specific categories of mental disorders that were organized under these six major headings. Most names of these specific categories would not be recognized by modern clinicians. The descriptions of these disorders, together with case histories that he included in this textbook do, however, suggest that many of the disorders he was discussing would have modern counterparts. For instance, under the heading "intellectual insanities," Hammond classified four disorders whose names seem odd by modern standards: intellectual monomania with exaltation, chronic intellectual monomania, reasoning mania, and intellectual subjective morbid impulses. In modern terms, these disorders probably would be called bipolar I disorder (manic episode), schizophrenia (continuous), narcissistic personality disorder, and obsessive compulsive disorder.

In Hammond's textbook, the lengthiest discussion was devoted to general paralysis, for which Hammond's name was dementia paralytica. As part of this discussion, Hammond included the proclamation by John Michler (quoted at the beginning of this chapter). In his discussion of general paralysis, Hammond emphasized the many medical symptoms associated with this disorder.

By the end of the nineteenth century, there were three broad theories about the etiology of this disorder. First, one school of thought believed that it was caused by alcoholism because the disorder primarily affected men, the age of onset was typically during the 30s and 40s (which is the same time of onset for severe alcoholism), and most men with dementia paralytica had substantial drinking histories. Second, was the theory that the disorder was caused by syphilis. Epidemiological surveys had found that over 80% of men with dementia paralytica had had syphilis. However, since no survey had documented 100% with a history of syphilis, many investigators suggested that syphilis was an important precondition to the development of dementia paralytica but was not the single cause. Hammond, for instance, was clear that syphilis was not the etiology because syphilis was associated with other forms of insanity. The third theory was more psychological in that the disorder was believed to be caused by moral depravity because persons who drank, who frequented prostitutes, and who were in the military were more likely to have the disorder. As additional evidence, dementia paralytica was known to be rare among priests and Quakers.

Research, attempting to provide evidence for or against these theories, was performed. For instance, a famous German psychiatrist named Kraft-Ebbing performed a study in which he injected serum from patients with syphilis into the blood streams of patients with dementia paralytica. Since it was known at the time that a person could only develop a syphilitic infection once, if any of the patients with dementia paralytica developed syphilis, it would prove that they had not had syphilis previously. Hence syphilis could not be the cause of the disorder. None of the 32 patients with dementia paralytica developed syphilis. Kraft-Ebbing concluded that syphilis was the cause of this disorder.

The conclusive evidence regarding the etiology of dementia paralytica occurred in the early twentieth century. In 1906, the bacillus that causes syphilis was isolated, and the Wasserman test for syphilis was developed. Plaut (1911) demonstrated that patients with dementia paralytica had positive Wasserman tests from blood samples and also from samples of cerebrospinal fluid. In 1913, two Americans, Noguchi and Moore, were able to identify the presence of the syphilitic bacillus in the brains of patients who had died from dementia paralytica (Quetel, 1990). The name of this disorder was changed again to reflect the new understanding of this disorder. It was called paresis or general paralysis associated with tertiary syphilis of the

central nervous system. However, even after the discovery of the cause of dementia paralytica, the development of antibiotics to treat the disorder was another 30 years in the future. Thus, many patients were treated by inoculating them with malaria (Braslow, 1995).

4.03.4 KRAEPELIN

Another important development at the turn of the century was the international focus on the classificatory ideas of a German psychiatrist named Emil Kraepelin (Berrios & Hauser, 1988). Kraepelin was a researcher who initially learned his approach to research in a laboratory organized by Wundt, one of the founders of modern experimental psychology. After completing his medical degree, Kraepelin became the medical director for an insane asylum in east Prussia. While in that setting, Kraepelin published a number of experimental psychology studies on persons with mental disorders. He also began to write textbooks about psychopathology for German medical students. Like most textbook authors, Kraepelin organized his chapters around the major categories of mental disorders that he recognized. The sixth edition of Kraepelin's textbook (Kraepelin, 1902/1896) attracted major international attention because of two chapters in these texts.

One chapter was about a disorder that Kraepelin described as dementia praecox (praecox = early) which was a form of psychosis that had a typical age of onset in adolescence. Kraepelin recognized three subtypes to this disorder: hebephrenic, catatonic, and paranoid. Kraepelin's chapter on dementia praecox paralleled the immediately preceding chapter on dementia paralytica. Just as dementia praecox had three descriptive subtypes, dementia paralytica also had three descriptive subtypes: a depressed form, a grandiose form, and a paranoid form.

The second chapter to attract attention was what Kraepelin named manic-depressive insanity. Kraepelin's observations of patients in asylums had led him to believe that the mania and depression (= melancholia) had the same type of course when these patients were observed over time. Both were episodic disorders. Moreover, nineteenth century clinicians had recognized that there were some patients who went from episodes of mania to episodes of depression and vice versa. These observations led Kraepelin to hypothesize that mania and depression were essentially flip sides of the same coin. Hence, he combined, what had been recognized since the times of the ancient Greeks as two disorders, into one mental disorder.

In 1917, the newly formed American Psychiatric Association (APA) adopted a classification system that was quite similar to the classification contained in Kraepelin's sixth edition of his textbook. This early twentieth century American classification included the concepts of dementia praecox and manic-depressive disorder. The classification also adopted the fundamental Kraepelinian distinction between the organic disorders, the functional psychoses, and the neurotic/character disorders.

In 1932, the APA officially adopted a new classification system as part of the Standard Classified Nomenclature of Diseases (APA, 1933). This new classification, however, did not attract much attention (Menninger, 1963).

4.03.5 EARLY DSMS AND ICDS

World War II led to a renewed emphasis on classification. During the war, nearly 10% of all military discharges were for psychiatric reasons. By the time the war ended, there were four major competing classification systems at use in the USA: (i) the standard system adopted by the APA in 1932; (ii) the US Army classification; (iii) the US Navy classification; and (iv) the Veterans Administration system (Raines, 1952). In response to this disorganization, the APA formed a task force to create a system that would become standard in the USA. The result was the *Diagnostic and statistical manual of mental disorders (DSM; APA, 1952)*. This classification is usually known by its abbreviated name: *DSM-I*. The *DSM-I* contained 108 different diagnostic categories.

The *DSM-I* was important for a number of reasons. First, the major rationale behind the *DSM-I* was to create a classification system that represented a consensus of contemporary thinking. Care was taken to include all diagnostic concepts that were popular in American psychiatry during the 1940s and 1950s. Thus, the *DSM-I* emphasized communication among professionals as the major purpose of classification and emphasized the need for psychiatric classification to be an accepted nomenclature that members of a profession can use to discuss their clinical cases. Consistent with this emphasis on communication, early versions of the *DSM-I* were revised, based on comments elicited in a questionnaire sent to 10% of the members of the APA. The *DSM-I* was finally adopted by a vote of the membership of the APA (Raines, 1952).

The emphasis on communication in the *DSM-I* led to a similar organizing movement at an international level. The international

psychiatric community had adopted a classification of mental disorders that was part of the *International statistical classification of disease, injuries and causes of death* (6th ed.) (*ICD-6* World Health Organization (WHO), 1948). The first *ICD* had been created in 1900 and was a medical classification of causes of death. The *ICD-6* was the first edition to include all diseases, whether they led to death or not. The classification of mental disorders in the *ICD-6* did not gain broad acceptance. A committee, chaired by the British psychiatrist Stengel, was formed to review the classification systems used by various countries and to make any necessary recommendation for changes to the WHO. What Stengel (1959) found was a hodgepodge of diagnostic systems between, and sometimes within, different countries. Stengel despaired over the confused state of international classification and said that the *ICD-6* did not serve as a useful nomenclature. A positive note in his review, however, concerned the *DSM-I*, which Stengel considered an advance over the other national classifications because of its emphasis on representing a well-organized nomenclature for a country.

As a result of Stengel's review, there was an international movement to create a consensual system that would be adopted by the WHO. The final product was the mental disorders section of the *ICD-8*. In the USA, the APA revised its *DSM* classification to correspond with the *ICD-8*. The US version of the *ICD-8* was known as the *DSM-II* (APA, 1968). The *DSM-II* had 185 categories. These categories were subdivided by a hierarchical organizational system. First, there was a distinction between psychotic and nonpsychotic disorders. The psychotic disorders were further subdivided into organic and nonorganic disorders. The classification of the organic disorders was in terms of etiology (e.g., tumors, infections, heredity, etc.). The nonorganic psychotic disorders primarily contained the Kraepelinian categories of schizophrenia and manic-depressive insanity. The nonpsychotic disorders were subdivided into eight subheadings including the neuroses (now called anxiety disorders), personality disorders, mental retardation, etc.

4.03.6 NEO-KRAEPELINIANS

After the publication of the *DSM-II*, psychiatric classification became a very unpopular topic. There were three general lines of criticism that were aimed at classification. First, the diagnosis of mental disorders was unreliable as shown by empirical research. Second, a number of critics attacked the implicit medical model

approach to psychopathology that was associated with the *DSM-I* and *DSM-II*. Third, sociologists became interested in a theory of labeling that suggested the process of classification was a process that stigmatized human beings who adopted unusual patterns of behavior and the act of diagnosis could lead to self-fulfilling prophecies.

The first of these criticisms was summarized in three different review articles by Kreitman (1961), Zubin (1967), and Spitzer and Fleiss (1974). All discussed various problems associated with the classification of mental disorders and why poor reliability was a significant issue. Zubin (1967) made an excellent case that the lack of uniform statistical procedures for estimating reliability was a serious methodological problem. Spitzer and Fleiss (1974) suggested a solution to this problem and showed how this solution could be applied retrospectively to earlier data. Kreitman (1961) probably had the most far-reaching analysis of the issue because he said that the unreliability problem had been overemphasized and that the more serious issue was the unexplored validity of diagnostic concepts.

The second issue criticism of the early *DSMs* and *ICDs* was the implicit acceptance of a medical model. Despite the dramatic etiological solution to dementia paralytica, most of the twentieth century research has been disappointing to those who believed that mental disorders are caused by underlying biological processes. Large amounts of research have attempted to discover the etiology of disorders such as schizophrenia, yet, despite interesting advances, a clear understanding of the cause of this disorder is not available. A psychiatrist named Thomas Szasz published a book titled *The myth of mental illness* (Szasz, 1961). He argued that mental disorders are not diseases, but instead are better conceptualized as "problems in living." He argued that psychiatrists had placed themselves into the role of moral policeman to control individuals with deviant behavior patterns. Szasz is now considered one of a group of critics of classification known by the title "antipsychiatrists." Others in this group are the British psychiatrist Laing (1967), the French psychoanalyst Lacan (1977), and recent authors such as Sarbin (1997), and Kirk and Kutichins (1992).

The third criticism of classification that became popular in the 1960s and 1970s was the labeling criticism. Sociologists such as Matza (1969) and Goffman (1961) suggested that the act of psychiatric diagnosis could lead to self-fulfilling prophecies in which the behavior of deviant individuals was constrained to become even more deviant. A dramatic demonstration of

the labeling criticism was contained in a controversial paper published by Rosenhan (1973). In this paper, Rosenhan and his colleagues gained admission to mental hospitals even though they reported everything about themselves factually except their names and one auditory hallucination. All but one of these pseudopatients were admitted with a diagnosis of schizophrenia and all were released with a diagnosis of schizophrenia in remission. The pseudopatients commented that most of the patients were aware that they did not belong there, even though the hospital staff never figured that out. In addition, the experiences of the pseudopatients supported the labeling concern. For instance, one pseudopatient reported being bored while being on a ward and walking around. A nurse noticed him pacing and asked if he was feeling anxious. Following the publication of the Rosenhan paper, an issue of the *Journal of Abnormal Psychology* in 1975 was devoted to commentaries on this controversial paper.

Partially in reaction to these criticism of classification, a new school of thought was formed in psychiatry called the neo-Kraepelinians (Klerman, 1978). This group of psychiatrists, initially an active collection of psychiatric researchers at Washington University in St. Louis, believed that psychiatry, with its psychoanalytic emphasis, had drifted too far from its roots in medicine. The neo-Kraepelinians emphasized that psychiatry should be concerned with medical diseases, that extensive research was needed on the biological bases of psychopathology, and that much more emphasis needed to be placed upon classification if knowledge about psychopathology was to grow. Klerman (1978) summarized the perspective implicit in the neo-Kraepelinian approach to psychiatry by listing the following tenets:

- (i) Psychiatry is a branch of medicine.
- (ii) Psychiatry should utilize modern scientific methodologies and base its practice on scientific knowledge.
- (iii) Psychiatry treats people who are sick and who require treatment.
- (iv) There is a boundary between the normal and the sick.
- (v) There are discrete mental illnesses. Mental illnesses are not myths. There is not one, but many mental illnesses. It is the task of scientific psychiatry, as of other medical specialties, to investigate the causes, diagnosis, and treatment of these mental illnesses.
- (vi) The focus of psychiatric physicians should be particularly on the biological aspects of mental illness.
- (vii) There should be an explicit and intentional concern with diagnosis and classification.

(viii) Diagnostic criteria should be codified, and a legitimate and valued area of research should be to validate such criteria by various techniques. Further, departments of psychiatry in medical schools should teach these criteria and not depreciate them, as has been the case for many years.

(ix) In research efforts directed at improving the reliability and validity of diagnosis and classification, statistical techniques should be utilized.

In 1972, a group of the psychiatric researchers at Washington University published a paper entitled "Diagnostic criteria for use in psychiatric research" (Feighner, Robins, Guze, Woodruff, Winokur, & Munoz, 1972). This paper listed 15 mental disorders that they believed had sufficient empirical evidence to establish their validity, and listed a set of diagnostic criteria for defining these disorders. The authors argued that a major problem in research about these disorders had stemmed from the lack of uniform definitions by different researchers of the disorders. They suggested that future research on any of these disorders should utilize the diagnostic criteria proposed in their paper.

The paper by Feighner et al. had a dramatic impact on American psychiatry. It was a heavily cited paper, probably the most frequently referenced journal article of the 1970s in psychiatry. The diagnostic criteria were immediately adopted and became the standard for psychiatric research. Moreover, the 15 categories described by Feighner et al. were expanded into a much larger set of categories, focusing primarily on the schizophrenic and affective disorders (Spitzer, Endicott, & Robins, 1975). This new classification was called the Research Diagnostic Criteria (RDC) and had an associated structured interview known as the SADS. Since the lead author of the RDC, Robert Spitzer, had been appointed as the new chairperson responsible for organizing the *DSM-III*, the RDC became the initial foundation from which the *DSM-III* developed.

4.03.7 *DSM-III, DSM-III-R, AND DSM-IV*

The *DSM-III* (APA, 1980) was a revolutionary classification. First, unlike the *DSM-I* and *DSM-II*, which had emphasized using consensus as the major organizing principle, the *DSM-III* attempted to be a classification based on scientific evidence rather than clinical consensus. For instance, the classification of depression was very different from the view of depression in the *DSM-I* and *DSM-II*, largely because of family history data gathered in research studies performed by the neo-Kraepelinians. In the earlier

DSMs, the primary separation of affective disorders was in terms of a psychotic vs. neurotic distinction. The *DSM-III* dropped this differentiation and, instead, emphasized a separation of bipolar vs. unipolar mood disorders. Second, the *DSM-III* discontinued the use of prose definitions of the mental disorders. The neo-Kraepelinians were impressed by the research data suggesting that the reliability of psychiatric classification, as represented in the *DSM-I* and *DSM-II*, was less than optimal (Spitzer & Fleiss, 1974). To try to help improve diagnostic reliability, virtually all mental disorders in the *DSM-III* were defined using diagnostic criteria stimulated by the innovative system used in the Feighner et al. paper. Third, the *DSM-III* was a multiaxial classification. Because the *DSM-I* and *DSM-II* were "committee products," the subsections of these classifications had different implicit organizing principles. For instance, in the *DSM-I/DSM-II*, the organic brain syndromes were organized by etiology, the psychotic disorders were organized by syndromes, and the neurotic disorders were organized according to ideas from psychoanalytic theory. In order to avoid the confusion inherent in the use of multiple organizing principles, the *DSM-III* adopted a multiaxial system that permitted the separate description of the patient's psychopathology syndrome (Axis I), personality style (Axis II), medical etiology (Axis III), environmental factors (Axis IV), and role disturbances (Axis V). The *DSM-III* was published in 1980 and contained 265 mental disorders. Moreover, the size of the manuscript for the *DSM-III* was 482 pages, a huge increase over the 92 pages of the *DSM-II*.

The revolutionary impact of the *DSM-III* led to changes in many areas of the mental health professions. One area of impact was in terms of research. As soon as versions of the *DSM-III* began to be disseminated to researchers interested in mental disorders, new studies began to appear that explored the adequacy of the diagnostic criteria used in this classification. The *DSM-III* was a marked stimulus for descriptive research in psychiatry and in the other mental health professions. Another area of impact was political. There was a major controversy that erupted in the late 1970s over the issue of whether the term "neurosis" should appear in the *DSM-III*. Spitzer and the neo-Kraepelinians had exorcized this term from the classification because of its psychoanalytic associations. The psychoanalysts lobbied intensely within the APA to have the term reintroduced. Although a compromise was achieved (Bayer, 1981), the psychoanalysts lost ground in this struggle, and their influence in organized psychiatry has continued to wane

since that time. A third area of impact for the *DSM-III* was economic. The *DSM-III* became very popular, sold well in the USA, and even became a surprisingly large seller to the international community as translations appeared. The sizeable revenues that accrued to the APA led to the formation of the American Psychiatric Press, which published subsequent versions of the *DSM* as well as many other books of clinical interest.

Despite its innovations and generally positive acceptance by mental health professionals, a number of criticisms were leveled at the *DSM-III*. One focus of criticism concerned the diagnostic criteria. Despite the intention to make decisions about the classificatory categories using scientific evidence, most diagnostic criteria were based on the intuitions of experts in the field. In addition, even though the goal when formulating diagnostic criteria was to make them as behavioral and explicit as possible, not all criteria met this goal. Consider part of the *DSM-III* criteria for histrionic personality disorder, for instance.

Characteristic disturbances in interpersonal relationships as indicated by at least two of the following:

- 1) perceived by others as shallow and lacking genuineness, even if superficially warm and charming
- 2) egocentric, self-indulgent and inconsiderate of others
- 3) vain and demanding
- 4) dependent, helpless, constantly seeking reassurance
- 5) prone to manipulative suicidal threats, gestures or attempts

Note that common language terms such as "shallow" were highly subjective. In addition, a criterion such as 3) requires an inference about motivations and reasons for behavior, rather than direct observation of behaviors. Finally, subsequent research showed that criterion 5) above actually was observed more frequently in borderline rather than histrionic patients (Stangl, Pfohl, Zimmerman, Bowers, & Corenthal, 1985).

A second major criticism of the *DSM-III* concerned the multiaxial system. First, diagnosing multiple axes required increased time and effort by clinicians, an exercise they were unlikely to do unless they were certain that the gain in information was significant. Second, the relative emphasis on these five axes in the *DSM-III* was sizeably different. In the *DSM-III* manual, almost 300 pages were devoted to defining the Axis I disorders, another 39 pages were spent on Axis II disorders, whereas only two pages each were devoted to Axes IV and V.

Moreover, Axes I and II were assessed using diagnostic categories, whereas Axes IV and V were measured using relatively crude, ordinal rating scales. Third, the choice of particular axes was also criticized. For instance Rutter, Shaffer, and Shepard (1975) had advocated the use of one axis for the clinical syndrome in childhood disorders with a second focusing on intellectual level. Instead, both clinical syndromes and mental retardation were Axis I disorders in the *DSM-III*. A group of psychoanalysts argued that defense mechanisms should have been included as a sixth axis. Psychiatric nurses advocated an additional axis for a nursing diagnosis, relevant to the level of care required by a patient.

Only seven years later, a revision to the *DSM-III* was published. This version, known as the *DSM-III-R* (APA, 1987), was intended primarily to update diagnostic criteria using the research that had been stimulated by the *DSM-III*. It was called a revision because the goal was to keep the changes small. However, the differences between the *DSM-III* and the *DSM-III-R* were substantial. Changes included renaming some categories (e.g., paranoid disorder was renamed delusional disorder), changes in specific criteria for disorders (e.g., the criteria for schizoaffective disorder), and reorganization of categories (e.g., panic disorder and agoraphobia were explicitly linked). In addition, six diagnostic categories originally in the *DSM-III* were deleted (e.g., egodystonic homosexuality and attention deficit disorder without hyperactivity) while a number of new specific disorders were added (e.g., body dysmorphic disorder and trichotillomania). As a result, the *DSM-III-R* contained 297 categories compared to the 264 categories in the *DSM-III*.

Associated with the *DSM-III-R* was the development of a major controversy that had political overtones. Among the changes proposed for the *DSM-III-R* was the addition of three new disorders: premenstrual syndrome (PMS), masochistic personality disorder and paraphilic rapism. These additions raised the ire of a number of groups, especially feminists. Concerning PMS, feminists argued that the inclusion of this disorder into the *DSM* would be the implicit assumption that the emotional state of women can be blamed on their biology. If it were to be a disorder, the feminists argued that PMS should be classified as a gynecological disorder rather than a psychiatric disorder. Masochistic personality disorder had been suggested for inclusion by psychoanalysts who pointed to the extensive literature on this category. Feminists, however, believed that this diagnosis would be assigned to women who had been physically or sexually abused. Thus, this

diagnosis would have the unfortunate consequence of blaming these women for their roles as victims. Finally, the proposal to include paraphilic rapism was also attacked. The critics argued that this diagnosis would allow chronic rapists to escape punishment for their crimes because their behaviors could be attributed to a mental disorder. Thus, these men would not be held responsible for their behaviors.

Because of the ensuing controversy, a compromise somewhat similar to the earlier compromises about homosexuality and neurosis was attempted. The authors of the *DSM-III-R* revised the names of the first two disorders (PMS and masochistic personality disorder) to perileteal phase dysphoric disorder and self-defeating personality disorder. They also deleted the proposal to add paraphilic rapism. In addition, another disorder, sadistic personality disorder, was added presumably to blame abuser as well as victims, thereby balancing the potential antifeminine connotations of self-defeating/masochistic personality disorder. This compromise was not successful. As a result, the executive committee for the American Psychiatric Classification decided not to include these categories in the body of the *DSM-III-R*. Instead, they were placed in an appendix as disorders needing more research (Walker, 1987).

The *DSM-IV* was published in 1994, contained 354 categories, and was 886 pages in length, a 60% increase over the *DSM-III* and almost seven times longer than the *DSM-II* (APA, 1994). There are 17 major categories in the *DSM-IV*:

- disorders usually first diagnosed in childhood
- cognitive disorders
- mental disorders due to a general medical condition
- substance-related disorders
- schizophrenia and other psychotic disorders
- mood disorders
- anxiety disorders
- somatoform disorders
- factitious disorders
- dissociative disorders
- sexual disorders
- eating disorders
- sleep disorders
- impulse control disorders
- adjustment disorders
- personality disorders
- other conditions that may be a focus of clinical attention

The *DSM-IV* retained a multiaxial system and recognized five axes (dimensions) along which patient conditions should be coded:

- Axis I clinical disorders
- Axis II personality disorders/mental retardation
- Axis III general medical conditions
- Axis IV psychosocial and environmental problems
- Axis V global assessment of functioning

As with the *DSM-III-R*, a major focus in the *DSM-IV* revision concerned diagnostic criteria. A total of 201 specific diagnoses in the *DSM-IV* were defined using diagnostic criteria. The average number of criteria per diagnosis was almost eight. Using this estimate, the *DSM-IV* contains slightly over 1500 diagnostic criteria for the 201 diagnoses.

To give the reader a glimpse of how the diagnostic criteria have changed from the *DSM-III* to the *DSM-IV*, the criteria for histrionic personality disorder are listed below:

A pervasive pattern of excessive emotionality and attention seeking, beginning by early adulthood and present in a variety of contexts, as indicated by five (or more) of the following:

- (1) is uncomfortable in situations in which he or she is not the center of attention
- (2) interaction with others is often characterized by inappropriate sexually seductive or provocative behavior
- (3) displays rapidly shifting and shallow expression of emotions
- (4) consistently uses physical appearance to draw attention to oneself
- (5) has a style of speech that is excessively impressionistic and lacking in detail
- (6) shows self-dramatization, theatricality, and exaggerated expression of emotion
- (7) is suggestible, i.e., easily influenced by others or circumstances
- (8) considers relationships to be more intimate than they actually are

In addition to presenting diagnostic criteria, the *DSM-IV* contains supplementary information about the mental disorders in its system. For instance, there are three pages of information about histrionic personality disorder including

diagnostic features (prose description of symptoms)
 associated features and disorders (mental disorders that are likely to co-occur)
 specific culture, age and gender features
 prevalence
 differential diagnosis (how to differentiate the disorder from others with which it is likely to be confused)

In order to help ensure that the *DSM-IV* would be the best possible classification of mental

disorders, the steering committee for this classification contained 27 members, including four psychologists. Reporting to this committee were 13 work groups composed of 5–16 members. Each work group had a large number of advisors (typically over 20 per work group).

There were three major steps associated with the activities of each work group. First, all work groups performed extensive literature reviews of the disorders under their responsibility. Many of these literature reviews were published in the journal literature. Second, the work groups solicited contributions of descriptive data from researchers in the field. Using these data, the work groups reanalyzed the data to decide which diagnostic criteria needed revision. Third, a series of field trials was performed on specific topics. For instance, the personality disorders work group performed a multicenter study on antisocial personality disorder which led to a significant alteration in the diagnostic criteria for that disorder.

The *DSM-IV* was not without controversy. For instance, the issues that had been raised in the *DSM-III-R* regarding premenstrual syndrome, masochistic personality disorder, and sadistic personality disorder continued in the *DSM-IV*. Interestingly, none of these three disorders were included in the *DSM-IV*. In fact, two (masochistic and sadistic personality disorders) completely disappeared from the classification. PMS remained in an appendix as a disorder “for further study.” Interestingly, 17 other disorders joined PMS in this appendix as did three possible new axes for the multiaxial system (defense mechanisms, interpersonal functioning, and occupational functioning).

Earlier editions of the *DSM* had few, if any, references to document the sources for any factual claims in these classifications. The *DSM-IV* attempt to overcome this problem was by publishing a five-volume companion set of sourcebooks. These sourcebooks are edited papers by members of the work groups. The intent of the sourcebooks is to provide a scholarly basis for understanding the specific decisions that were made by the work groups.

4.03.8 ICD-9, ICD-9-CM, AND ICD-10

Earlier in this chapter, the point was made that the *DSM-II* and *ICD-8* were virtually identical because the American psychiatric community had joined an international movement to create a consensual classification. With the revolutionary *DSM-III*, American psychiatry reversed itself and created a radically new classification based upon the purpose of

description, rather than emphasizing a system that would be acceptable world-wide.

The editions of the *ICDs* were intended to be revised every 10 years. The *ICD-8* was published in 1966; the *ICD-9* came out in 1977. The mental disorders section of the *ICD-9* was very similar to the *ICD-8/DSM-II* (WHO, 1978). The psychotic/nonpsychotic distinction was the primary hierarchical distinction among categories. The psychotic disorders were further subdivided into organic and functional psychoses. There were 215 categories in this system, and the *ICD-9* was published in a monograph that was 95 pages in length.

The USA has signed an international treaty that obliges it to use the *ICD* as the official medical classification. Thus, when the *DSM-III* was created, an odd numeric coding scheme was incorporated so that the *DSM-III* categories could be incorporated with the *ICD-9* framework. To understand this, below is an overview of the specific diagnostic categories under the general heading of anxiety disorders in the *DSM-III*:

Phobic disorders

- 300.21 Agoraphobia with panic attacks
- 300.22 Agoraphobia without panic attacks
- 300.23 Social phobia
- 300.29 Simple phobia

Anxiety states

- 300.01 Panic disorder
- 300.02 Generalized anxiety disorder
- 300.30 Obsessive compulsive disorder
- Post-traumatic stress disorder
- 308.30 Acute
- 309.81 Chronic or delayed
- 300.00 Atypical anxiety disorder

Notice that the coding scheme for the anxiety disorders in the *DSM-III* is not what one might expect. Most of the anxiety disorders are coded with 300.xx numbers. However, the two forms of post-traumatic stress disorder are coded 308.30 and 309.81. Notice also that the first number after the decimal point is somewhat irregular. The phobic disorders, listed first in the *DSM-III*, are given 300.2x codes whereas the anxiety states are coded 300.0x or 300.30.

To understand why the *DSM-III* codes appear this way, below is a listing of the specific neurotic disorders in the *ICD-9*:

Mental disorders (290–319)

Nonpsychotic mental disorders (300–316)

Neurotic disorders (300)

- 300.0 Anxiety states
- 300.1 Hysteria
- 300.2 Phobic states
- 300.3 Obsessive-compulsive disorders

- 300.4 Neurotic depression
- 300.5 Neurasthenia
- 300.6 Depersonalization syndrome
- 300.7 Hypochondriasis
- 300.8 Other neurotic disorders
- 300.9 Unspecified

In the *ICD-9* system, all diagnoses have four-digit codes. The codes for all mental disorders range from 290 to 319. The 29x disorders are the psychotic disorders; 300–315 are reserved for nonpsychotic disorders; and 316–318 are codes for classifying mental retardation. The first subheading under the nonpsychotic disorder is the neurotic disorders. Notice that this subheading includes what the *DSM-III* recognizes as anxiety disorders but it also includes categories that the *DSM-III* placed under other headings (e.g., neurotic depression = dysthymic disorder and depersonalization syndrome). Because the *DSM-III* anxiety disorders were mostly found under the *ICD-9* neurotic disorders, these categories were given 300.xx codes. However, post-traumatic stress disorder (chronic; PTSD) was given a code number of 309.89 because it was included under the general *ICD-9* heading of adjustment reactions. Notice also that PTSD has an xxx.8 coding. In the *ICD-9* coding system, all diagnoses with an xxx.8 code represent country-specific categories that are generally recognized by the international psychiatric community. Thus, PTSD (309.89) is a US category that has no clear equivalent in the international diagnostic system. Another *DSM-III* category with a coding that reflects a similar status is borderline personality disorder (301.83).

In order to blend the *DSM-III* with the *ICD-9* so that a consistent coding system would be used, a US version of the *ICD-9* was created. This new version was the *ICD-9-CM* (where CM stands for “clinical manual”). The *ICD-9-CM* is the official coding system that all physicians and mental health professionals must use when assigning diagnostic codes. However, American clinicians do not need to refer to the *ICD-9-CM* because the applicable codes are listed in the printed versions of the *DSM-III*, *DSM-III-R* and *DSM-IV*.

As noted earlier, the *DSM-III* and its successors (*DSM-III-R* and *DSM-IV*) were resounding successes. Not only did these systems become dominant in the USA, but they also achieved substantial popularity among European mental health professionals (Mezzich, Fabrega, Mezzich, & Coffman, 1985). The proponents of the *ICD* were somewhat resentful (Kendell, 1991). Thus, when the *ICD-10* was created, substantial changes were made that utilized innovations from the *DSM-III*.

First, like the *DSM-III*, the *ICD-10* went through extensive field testing (Regier, Klaeber, Roper, Rae, & Sartorius, 1994). There were two major goals in the field testing: (i) to ensure that the *ICD-10* could be used in a reliable way across clinicians, and (ii) to examine the acceptability of the mental disorder categories contained in this system. The data reported regarding both of these goals have given a favorable view of the *ICD-10*.

The second important innovation of the *ICD-10* is that the mental disorders section is published in two forms. One form, the blue manual, is subtitled *Clinical descriptions and diagnostic guidelines* (WHO, 1992). The blue manual contains prose definitions of categories and is primarily intended for clinical use. The other, the green manual, is like the *DSM-III* in that the categories are defined using explicit diagnostic criteria with rules regarding how many criteria must be met in order for a diagnosis to be made (WHO, 1993). The green manual is intended for research use.

The complete *ICD-10* is organized into a series of 21 chapters, one of which is Chapter V (labeled with the prefix F) about "Mental and behavioural disorders." Other chapters in the *ICD-10* are:

Chapter I A–B Infections and parasitic diseases

Chapter II C–D Neoplasms

Chapter X J Diseases of the respiratory system

Chapter XXI Z Factors influencing health status and contact with health services

In terms of classificatory size, the mental disorders section of the *ICD-10* and the *DSM-IV* are reasonably similar. The *DSM-IV* contains 354 categories organized under 17 major headings. The *ICD-10* has 389 categories that are structured under 10 major headings. One ironic feature of the *ICD-10* is that it did not adopt a multiaxial system of classification. This decision is ironic because the idea originated in Europe, most prominently by a Swedish psychiatrist, Essen-Møller (1971).

4.03.9 CONTROVERSIES AND ISSUES

Although the *DSM-III* and its successors are usually viewed as substantial improvements in the classification of mental disorders, a number of controversies and issues still remain concerning the classification of mental disorders. The remainder of this chapter attempts to provide an overview of some of these issues.

4.03.9.1 Organizational Models of Classification

There are four organizational models of classification that have been frequently discussed in the literature. Often these models are seen as competing and incompatible. For instance, most discussions of the first two models, the categorical and the dimensional, are presented as if the mental health professions must choose between them. However, hybrid models are possible (Skinner, 1986) and perhaps even the most probable.

Mental disorders are usually discussed as if they are categories. Thus, a categorical model is often implicitly assumed to be the structural model for psychopathology (Kendell, 1975). The tenets of this model are listed below:

- 1.1 The basic objects of a psychiatric classification are patients
- 1.2 Categories should be discrete, in the sense that the conditions for membership should be able to be clearly ascertained
- 1.3 Patients either belong or do not belong to specified classes (categories)
- 1.4 The members of a category should be relatively homogeneous
- 1.5 Categories may or may not overlap
- 1.6 In the borderline areas where categories may overlap, the number of overlapping patients should be relatively small
- 1.7 Cluster analytic methods can be used to initially identify categories
Discriminant analysis is used to validate categories. (Blashfield, 1991, p. 14)

The *DSMs*, particularly the *DSM-III* and its successors, are seen as fitting a categorical model. According to the categorical model, the unit of analysis is the patient. Diagnoses refer to classes of patients. Patients either are or are not the members of the categories. The categorical model assumes that some type of definitional rule exists by which the membership in a category can be determined. Moreover, membership in a category is considered to be a discrete, all-or-nothing event. An animal is either a cat or not a cat. A patient is either a schizophrenic or not a schizophrenic.

An important assumption of the categorical model is that members of a category are relatively homogeneous. All animals that belong to the class of "birds" should be reasonably similar morphologically. This is not to say that all birds must be alike. Certainly a robin and sparrow have a number of obvious differences. Yet they are more like each other than either is to a lynx, for instance. In the same way, two schizophrenic patients should be relatively similar. Both may have different symptom

pictures, but their symptom pictures should be more similar to each other than either is to an antisocial patient (Lorr, 1966).

Classes in a categorical model may or may not overlap. Most uses of the categorical model typically assume that nonoverlapping categories are an ideal condition, but recognize that this condition may not always happen. Thus, overlap among categories is often treated like measurement error—a condition to be tolerated, but minimized. However, there are categorical models that have been developed in which the categories are assumed to overlap (Jardine & Sibson, 1971; Zadeh, 1965). In these models, overlap is not error. Categories are fuzzy sets whose boundaries of membership do not need to result in mutually exclusive groupings.

According to the assumption of relative homogeneity, the number of patients who clearly belong to one and only one category should be relatively frequent, whereas patients who fall in the overlapping, borderline areas between categories should be relatively infrequent. A sparrow-lynx should not occur if categories are to have the necessary homogeneity that allows them to be separable constructs (Grinker, Werble, & Drye, 1968).

Finally, the methods that have been developed to find the boundaries among categories are called cluster analytic methods (Everitt, 1974). Generally, these methods analyze a large matrix of descriptive data on patients, and attempt to form clusters (categories) of relatively homogeneous patients in terms of the descriptive variables that were gathered on the patients. Cluster analysis was used in the 1960s and 1970s to create new descriptive classifications. In the last decade, most researchers have abandoned the use of these methods because of pragmatic difficulties in their application and because of unsolved statistical issues. Meehl, however, has developed a related method for isolating categories that he believes has promise (Meehl, 1995).

Although the categorical model, as presented above, seems to be a common sense model of psychiatric classification, the recent *DSMs* clearly do not adhere to this model. First, as noted above, a categorical model assumes that the unit of analysis is the patient and that groups of patients are members of similar sets called mental disorders. The authors of the *DSM-III*, *DSM-III-R* and *DSM-IV* explicitly reject this approach. They state that these classifications are not intended to classify individual patients (Spitzer & Williams, 1987). Instead, these recent *DSMs* state that they are classifying disorders (rather than patients).

A second structural model is the dimensional model. The tenets for this model are:

- 2.1 The basic unit of the dimensional model is a descriptive variable (e.g., a symptom, a scale from a self-report test, a laboratory value, etc.)
- 2.2 Dimensions refer to higher-order, abstract variables
- 2.3 A dimension refers to a set of correlated descriptive variables
- 2.4 There are a relatively small number of dimensions compared to the number of descriptive variables, yet the dimensions account for almost as much reliable variance as do the larger number of descriptive variables
- 2.5 Dimensions themselves may be correlated or independent
- 2.6 The methods used to identify dimensions are exploratory factor analysis and multidimensional scaling. Confirmatory factor analysis can be used to test a specific dimensional model. (Blashfield, 1991, p. 15)

For the dimensional model, the basic units of analysis are the descriptive variables. Thus, the dimensional model focuses on symptoms, behaviors, diagnostic criteria, scales from self-report tests, and the like. The dimensional model summarizes these descriptive variables by forming higher-order abstract variables that can serve to represent the original measurement variables. Each of these higher-order abstract variables constitutes a dimension through its conceptualization as occurring on a continua. Patients can have scores anywhere along these dimensions.

A major test of a dimensional model is parsimony. The specific dimensions in such a system should account for most of the systematic, reliable variance that exists within the original set of descriptive variables. If the dimensions do not account for the reliable variance in the original descriptive variables, then using the dimensions will sacrifice a great deal of information and the original variables should be used rather than the smaller set of dimensions.

A third structural model that is often discussed regarding the classification of mental disorders is a disease model. The basic assumption in this model is that all diagnostic categories refer to medical diseases (Wing, 1978). In effect, this model is a modern extension of Griesinger's famous nineteenth-century dictum that all mental disorders are diseases of the brain (Stromgren, 1991). The tenets of the disease model are:

- 3.1 The fundamental units are biological diseases of individual patients (essentialism)
- 3.2 Each diagnosis refers to a discrete disease
- 3.3 Diagnostic algorithms specify objective rules for combining symptoms to reach a diagnosis
- 3.4 Adequate reliability is necessary before any type of validity can be established

- 3.5 Good internal validity will show that the category refers clearly described patterns of symptoms
- 3.6 Good external validity will mean that the diagnosis can be used to predict prognosis, course and treatment response. (Blashfield, 1991, p. 8)

Some authors have assumed that a categorical model and a disease model are the same. These models are not identical. A categorical is neutral about the existential status of the categories in its system. A disease model adopts a stronger view. Diseases do have an existential status. Diseases are real. The goal of medical research is to identify, describe, understand and eventually treat these diseases. This belief in the reality of diseases is associated with a broader view about the status of scientific concepts known as essentialism.

Notice also that diseases are not necessarily categorical, at least as this model was described above. For instance, more than one disease can occur in the same patient. In fact, some diseases are very likely to co-occur (e.g., certain sarcomas have high frequency in patients with AIDS). Thus, diseases do not refer to mutually exclusive sets. In addition, there are diseases that are conceptualized as dimensional constructs. Hypertension is the most common example. Patients with hypertension vary along a continuum. A categorical scaling of this continuum is possible, but imposing a categorical separation on this continuum is arbitrary.

The fourth model of classification is the prototype model. Cantor, Smith, French, and Mezzich (1980) have suggested that this model is superior to the implicit categorical model of psychiatric classification. For those readers who do not know what the prototype model is, the easiest way to conceptualize this model is through an example.

According to the prototype model, if a mother wanted to teach a child what "anger" means, she would not say "Steven, you need to understand that anger is an emotion that many of us feel when we are frustrated or upset." Instead, at some point when little Steven is upset because he has to go to bed and he tries to hit his mother, she would say "I know that you are angry, Steven, but it is time to go to bed." And on another day, when Steven is upset because another child took one of his toys, his mother might say, "You are feeling angry, Steven. Being angry is natural, but you should not hit the other child. Maybe if you ask Carol she will return your toy."

In effect, a child learns a concept by being presented with instances of the concept. Once the child is able to associate the instances with a verbal label (i.e., the word "angry"), then the

child will have a grasp of the concept. Later, the child will learn to abstract the essential features of the concept. This occurs by making observations about similarities that occur among instances of the concept (e.g., internal feelings, interpersonal context, etc.). Russell and Fehr (1994) provide an interesting and more complete discussion of the concept of anger from a prototype perspective.

Another important aspect of the prototype model is the idea that not all instances of a concept are equally good representatives of the concept. A robin, for instance, is a good exemplar of a bird. Robins are about the same size as most birds; they have feathers; they can fly; etc. Penguins, however, are not a good exemplar. Penguins are larger than most birds; they cannot fly; they do have feathers, although, to a child, their covering probably seems more like fur than feathers; etc.

The above presentation of the prototype model is easy to understand and seems like a common-sense view of classification. However, advocates of the prototype model argue that this model is radically different than a categorical model (Barasalou, 1992; Russell, 1991). According to the categorical model, classificatory concepts are defined by listing the features that are sufficient for making a diagnosis. If a given instance has a sufficient number of these features, then that instance is a member of the classificatory concept. Moreover, all members of a concept are equal. A square is a square. One square is not squarer than another square. In contrast, the prototype model does stipulate that some instances of a concept are better exemplars than others. The Glenn Close character in *Fatal attraction* is a better representation of borderline personality disorder than the Jessica Walter character in *Play Misty for me*.

The basic tenets of the prototype model are presented below:

- 4.1 Diagnoses are concepts that mental health professionals use (nominalism)
- 4.2 Categories are not discrete
- 4.3 There is a graded membership among different instances of a concept
- 4.4 Categories are defined by exemplars
- 4.5 Features (symptoms) are neither necessary nor sufficient to determine membership
- 4.6 Membership in a category is correlated with number of features (symptoms) that a patient has. (Blashfield, 1991, p. 11)

The major difference between the disease and the prototype model is that the latter is associated with nominalism. Nominalism is the position that the names of diseases are just convenient fictions that clinicians use to organize information. Diagnostic terms do not have

some underlying reality of their own. Diagnostic concepts are names. These concepts exist simply to meet some functional goal.

The preceding discussion of organizational models of the classification of psychopathology is overly simplistic. Each of these models, when examined more closely, can become quite complex, and the apparent distinctions between the models blur. Two instances of this complexity will be described. First, although the categorical and dimensional are usually presented as if they are competing and antagonistic, Skinner (1986) has suggested that these models are actually complementary. He suggested that the measurement model associated with a dimensional perspective is the more fundamental of the two models. The dimensional model only assumes that, in order to assess a patient, a clinician should gather information on specific descriptive variables that are correlated and that can be summarized by higher-order variables (dimensions). The categorical model also assumes that descriptive variables can be sorted into dimensions. But, in addition, the categorical model asserts that the patients themselves "cluster" into groups that are relatively homogeneous on these dimensions. Thus, from Skinner's hybrid perspective, a pure categorical model makes stronger assumptions about descriptive data than does a dimensional model. However, psychological models of human social cognition suggest that categorical models are more basic than are dimensional models (Wyer & Srull, 1989).

A second example of the complexity of these models is associated with Barsalou's distinction between "prototype" and "exemplar" models. There are two types of approach that can be used to define a concept: intensional definitions and extensional definitions. An intensional definition lists the features that can be used to separate the concept from related concepts (e.g., a square is a four-sided closed figure whose sides are equally long and occur at right angles to each other). In contrast, an extensional definition is a definition by listing the members of the category (e.g., the 1957 New York Yankees included Roger Marris, Yogi Berra, Micky Mantle, etc.). Barsalou says that a prototype model uses an intensional definition for categories in which the prototype represents the average (centroid) of the concept. Using the example of a child learning about birds, Barsalou suggests that the reason that a robin is a prototype for bird, whereas a penguin is not, is that robins have the average features of most birds (small size, bright coloring, migration, food choices, etc.). Penguins, in contrast, are statistically unusual on these dimensions. An exemplar model, according to Barsalou, uses a

particular type of extensional definition in which a concept is defined by an outstanding or exemplary instance of the concept. Thus, Micky Mantle might be a good exemplar of the 1957 New York Yankees even though Mantle's batting prowess was hardly average. In the same way, Abraham Lincoln might be seen as exemplar of American presidents, even though he was not average for this set of individuals.

4.03.9.2 Concept of Disease

The discussion of the disease model vs. the prototype model led to a brief introduction regarding the dualism of essentialism vs. nominalism. This dualism is associated with a complicated problem that has bothered writers about classification throughout the last two centuries. What is a disease? Do the two concepts of "mental disease" and "mental disorder" have the same meaning? To discuss the issues associated with the meaning of "disease," the writings of a British internist named Scadding will be discussed. At the end of this section, other approaches to the concepts of "disease" and "disorder" are briefly introduced.

Scadding's (1959) first attempt to discuss the meaning of disease occurred in a short essay. This essay offered his first general definition of disease which read:

The term "a disease" refers to those abnormal phenomena which are common to a group of living organisms with disturbed structure or function, the group being defined in the same way.

In effect, Scadding was saying that a disease was associated with a cluster of signs and symptoms (i.e., abnormal phenomena) that are associated with some functional or structural disturbance in the human body. Scadding went on to argue that a disease had (i) defining characteristics and (ii) diagnostic criteria. The defining characteristics refer to the indications that prove the presence of the disease (e.g., locating syphilitic bacilli in the brains of individuals with paresis). In contrast, the diagnostic criteria are signs and symptoms of the disease that may or may not be present (e.g., motor paralysis, grandiose delusions, and sluggish pupillary response to light would be possible diagnostic criteria for paresis).

Ten years later, Scadding (1969) revised his definition of disease to read as follows:

A disease is the sum of the abnormal phenomena displayed by a group of living organisms in association with a specified common characteristic or set of characteristics by which they differ from the norm for their species in such a way as to place them at biological disadvantage.

There are four important points to note about this second definition of disease. First, the emphasis is on abnormal phenomena. Scadding wanted to be quite clear that the name of the disease does not refer to the etiologic agent causing the disease. That is, tuberculosis is not simply defined by the presence of a particular bacterium, *Mycobacterium tuberculosis*. To have tuberculosis a patient must manifest the symptoms of the disease as well as the anatomical changes (i.e., the formation of characteristic lesions called tubercles in the lung) associated with this disease. This distinction is important because there are other bacilli, besides *Mycobacterium tuberculosis*, which can cause these lesions and the symptom pattern of tuberculosis.

Second, the definition contains the rather vague phrase "common characteristic." Scadding argued that there are three general ways of characterizing any individual disease: (i) a clinical-descriptive approach, (ii) a morbid anatomical approach, and (iii) an etiological approach. (Note that these approaches were presented almost a century earlier by Hammond (1883), as previously discussed.) The clinical-descriptive approach is simply the description of the "syndrome." That is, the clinical-descriptive approach outlines a loose cluster of signs and symptoms that are correlated in their appearance in patients. For instance, the clinical-descriptive approach to defining diabetes focuses on frequent urination, an intense thirst, and rapid loss of weight as indications of this disorder. The clinical-descriptive approach dominated when the *DSM-III* and its successors were created. The second approach concerns morbid anatomy. This refers to the anatomical changes in the body's structure associated with the disease. For diabetes mellitus, a morbid anatomy view might define this disease in terms of the destruction of β -insulin-producing cells in the pancreas. Finally, the etiological approach would be to define a disease in terms of the syndrome caused by a known and specifiable etiological process. For Type I diabetes mellitus, this might be an autoimmune process whose exact details have yet to be specified. For paresis, the etiological agent is the effect of the syphilitic bacillus on the central nervous system of the affected individual.

Scadding commented that, historically, knowledge about diseases typically proceeds from clinical-description to morbid anatomy to etiology. Certainly his observation seems to be correct when applied to the history of paresis. He argued that any of these approaches to characterizing a disease are appropriate. That is, a disease can be defined in terms of a clinical syndrome; or it can be defined by some

associated morbid anatomy; or it can be defined through a recognition of its etiological cause.

The third point to note about Scadding's definition is its emphasis on norms, in that disease refers to an abnormality. To be a disease, the condition must refer to phenomena that are statistically deviant. For instance, most of us have various types of bacteria that normally reside in our intestines and which are important in the digestive process. The presence of these bacteria do not define a disease. The effects of these bacteria are normative. In fact, individuals without these bacteria have abnormal digestive processes.

Finally, the definition ends with the term "biological disadvantage." Scadding introduced this term because he recognized that not all nonnormative aspects of human structure and functioning should be called diseases. For instance, some individuals produce an abnormal amount of ear wax. However, this should not define a disease unless there is some biological disadvantage associated with this condition. Although the term biological disadvantage is not more precisely specified, its general meaning seems clear: syphilis and diabetes place an individual at biological disadvantage since both can lead to death if untreated.

In 1979, Scadding and two Canadian authors (Campbell, Scadding and Roberts, 1979) extended their ideas about disease by studying what physicians and nonphysicians meant by the concept of disease. They published a survey that they had conducted regarding the meaning of disease. To conduct their survey, these authors read a list of possible diseases to four groups of individuals: (i) a group of medical faculty, (ii) a group of nonmedical faculty, (iii) a sample of general practice physicians, and (iv) a sample of youth in British and Canadian schools. The subjects in this study were asked to note whether the terms being read aloud referred to diseases or not. In addition, the subjects were asked to assign degree of confidence ratings to their decisions.

At the top of the list of conditions that were viewed as diseases are infections (malaria, tuberculosis, syphilis, measles). Virtually everyone in the four groups, whether physicians or nonphysicians, agreed that these terms referred to diseases. Syphilis, for instance, was considered a disease by over 90% of the subjects in all groups.

At the bottom of the list were concepts that were not considered diseases by these subjects. Two terms that were seen as referring to diseases by less than 30% of all four groups were drowning and starvation. Many of the terms at the bottom of Scadding's list might be described

as injuries, i.e., traumas that affect bodily functioning and that were caused by identifiable external events such as a car accident (e.g., skull fracture) or ingestion of a substance (e.g. barbituate overdose, poisoning).

The psychiatric concepts in the list (schizophrenia, depression, and alcoholism) were ranked in the middle. There was considerable variance among the four groups regarding whether these concepts referred to diseases. For instance, faculty of medical schools rated these three concepts in the following order: schizophrenia (78%), depression (65%), and alcoholism (62%). Children in secondary schools had quite different impressions of what is considered to be a disease: alcoholism (76%), schizophrenia (51%) and depression (23%).

One factor that had a large influence on whether a term referred to a disease concerned the role of a physician in the diagnosis or treatment of the disorder. Malaria and syphilis require a doctor to diagnose and treat. In contrast, starvation can be identified and treated by nonmedical individuals. The latter is also true of acne and hemorrhoids, although the intervention of physicians can prove useful for both. Consistent with this view, acne and hemorrhoids were ranked in the middle of the list. The potential role of nonphysicians in the treatment of mental disorders may also account for the occurrence of schizophrenia, depression and alcoholism in the middle of the same list. Scadding et al (1979, p. 760) concluded their paper with the following interesting comment:

Most people without medical training seem to think of a disease as an agent causing illness. The common concept of "disease" is essentialist: diseases exist, each causing a particular sort of illness. Doctors tend to adopt a more nominalist position, but they obviously retain remnants of belief in the real existence of diseases.

When viewed from this dualism of an essentialist vs. a nominalist perspective, Scadding had started his definitional attempts from an essentialist perspective but, by the time of his last writings on the topic, he was suggesting that a nominalist view was preferred. Interestingly, the writings of a prominent British psychiatrist, Kendell, who has also tried to solve this issue, have followed the same progression. His ideas shifted from a paper trying to settle on an essentialist meaning of disease (Kendell, 1976) to a skeptical discussion of how the disease model fails to explain alcoholism (Kendell, 1979) to a nominalist view (Kendell, 1986).

Because this nominalism vs. essentialism dualism is so important, the approach of Wulff, Pedersen, and Rosenberg (1986) is discussed

briefly. To understand the dualism, Wulff et al. discussed possible ways to classify defective grandfather clocks as an example.

Suppose that one examines how people who work in a repair shop might classify clocks. The receptionist, who knows very little about the workings of the clocks, might classify them descriptively. Thus, some clocks would be placed together because they do not work after being wound; others have broken faces, arms or other parts; and still others do not keep time accurately. Another person who might classify the grandfather clocks would be the bookkeeper of the shop. This individual might classify the clocks according to the manufacturer and cost of the clock. A third person who might classify the clocks is the repairman. He might organize clocks anatomically into those with accumulated dirt impeding their normal functioning, those needing replacement parts, and those with weighting mechanisms that have become unbalanced. Finally, the owner of the repair shop, when reporting back to various manufacturers about the causes of clock malfunctions, might classify the clocks etiologically. That is, she might report about clocks that have had little care, clocks that become worn over various time intervals of ownership, and clocks that developed problems after being moved or damaged.

Which of these classifications—descriptive, cost oriented, anatomical, or etiological—is the true or best classification of defective grandfather clocks? From the nominalist perspective, none of these classifications is inherently the best. Each of these classifications is imposed by the needs of the particular individual using the classification. Each classification serves a function. For any particular function, one classificatory system may be preferable. But none of these is the true classification of defective clocks.

Notice that this apocryphal classification of defective clocks is analogous to the approaches to defining disease suggested by Scadding: clinical-descriptive, morbid anatomical and etiological. The cost oriented classification was simply added as an analogy to how medical classifications are used by the insurance industry in the USA.

Wulff's defective clock analogy was borrowed from the British philosopher, John Locke. Locke had argued that classificatory systems are inherently nominalist, even though the ultimate goal is often essentialist:

Therefore we in vain pretend to rank things in sorts, and dispose them into certain classes, under names, by their *real essences*, that are so far from our discovery or comprehension. (Wulff et al., 1986, p. 75)

In this regard, it is interesting to contrast Locke's view of classification to those of his physician friend, Thomas Sydenham. Believing in an essentialist view of disease, Sydenham made the following statement which has been quoted repeatedly since then:

Nature, in the production of disease, is uniform and consistent . . . The selfsame phenomena that you observe in the sickness of a Socrates you would observe in the sickness of a simpleton.

In other words, diseases do exist. They do have an essence. It is the business of medical research to discover what these essences are. It was the belief in this essentialist perspective that led nineteenth century researchers to solve the etiological issues associated with dementia paralytica.

Scadding and Wulff et al. warned about the dangers of essentialist thinking when applied to disease. For instance, Scadding noted that often we mistake the disease for the cause of the disease. Noguchi and Moore (Quetel, 1990) discovered the syphilitic bacilli in the brains of individuals with paresis. Hence, we might say that paresis occurs when syphilis invades the central nervous system. But the last sentence is misleading. Syphilis is not an organism. The bacteria, *Treponema pallidum*, is an organism and it could be said to invade the central nervous system. But even if this bacteria were present in the brain of an individual, that presence does not mean that the individual has paresis. To have paresis the individual must manifest the characteristic symptoms and anatomical changes associated with paresis.

Wulff et al. end their discussion of nominalism vs. essentialism with the following statement:

The philosophical problem which underlies the discussion in this chapter is the age-old dispute about universals, and we have tried to navigate between the Scylla of essentialism (or Platonism) and the Charybdis of extreme nominalism. Essentialism underlines correctly that any classification of natural phenomena must reflect the realities of nature, but it ignores the fact that classifications also depend on our choice of criteria and that this choice reflects our practical interests and the extent of our knowledge. Nominalism, on the other hand, stresses correctly the human factor, but the extreme nominalist overlooks that classifications are not arbitrary but must be molded on reality as it is. (Wulff et al., 1986, pp. 87–88)

As mentioned earlier, defining the concepts of “disease” and/or “disorder” raise complicated issues, and the preceding discussion does not adequately cover the literature. One American

author, in particular, has attracted substantial attention in the 1990s for his writings on this issue. Wakefield (1992, 1993) initially addressed this issue by providing a detailed critique of the definition of mental disorder that appeared in the *DSM-III*. Following this seminal paper, other theoretical papers (Wakefield, 1997a, 1997b) proposed a “harmful dysfunction” view of how to define mental disorders. A special section of the *Journal of Abnormal Psychology* has been devoted to a discussion of Wakefield's ideas.

Besides Wakefield's writings, there are other important discussions of this definitional issue including an overview by Reznick (1987), a book by Wing (1978), and an edited book on philosophical issues associated with classification (Sadler, Wiggins & Schwartz, 1994).

4.03.9.3 Two Views of a Hierarchical System of Diagnostic Concepts

Categories in the classification of mental disorders are organized hierarchically. This structural arrangement is commonly recognized but, since the publication of the *DSM-III*, two different views about this hierarchical structure have been discussed. Since these two views are often confused, the next section briefly discussed them.

The first approach to the meaning of hierarchy is nested set approach. Consider, for instance, the *DSM-II* classification of mental disorders. In this system, there are two broad categories of disorders: (I) psychotic disorders and (II) nonpsychotic disorders. The psychotic disorders are further subdivided into (I.A) the organic disorders and (I.B) the nonorganic disorders. The nonorganic psychotic disorders in the *DSM-II* were subdivided into three categories: (I.B.1) schizophrenic disorders, (I.B.2) major affective disorders, and (I.B.3) paranoid disorders. Then the schizophrenic disorders were subdivided into various subtypes: (I.B.1.a) simple type, (I.B.1.b) hebephrenic type, (I.B.1.c) catatonic type, and so on. Notice that this organization of mental disorders has a similar outline to the organization of categories in the biological classification. Any patient, for instance, who would be diagnosed as being hebephrenic, would also be considered as having schizophrenia disorder (the next highest, inclusive category) as well as having a psychotic disorder (an even higher level, inclusive category). Thus, this approach to hierarchy is called a nested set approach because the categories low in the system refer to sets of patients that are included (nested) in higher order categories. This approach parallels the

classification of biological organisms in which any animal who is a member of the species *Felis catus* (housecat) is also member of the genus *Felis* (cats) and a member of an even higher order category of *Mammalia* (mammal).

The other approach to hierarchy is called a pecking order view. This view can be best understood by making analogy to the hierarchical organization of rank among military officers. A colonel is higher in rank than a major who in turn is higher in rank than a lieutenant. In this pecking order structure, a colonel can give orders to a lieutenant, but a lieutenant cannot issue orders to a colonel. Thus, the pecking order in military rank concerns lines of authority. A colonel has authority over a lieutenant. Notice, however, that there is no membership nesting in these categories. If a particular individual is a lieutenant, then that individual is not a colonel even though a colonel is higher in the hierarchy than a lieutenant.

To understand how this analogy to the hierarchical arrangement of military rank can be applied to psychiatric classification, consider the following order of general mental disorders:

- organic mental disorders
- schizophrenic disorders
- affective disorders
- anxiety disorders
- personality disorders.

In terms of the pecking order meaning of hierarchy, this order means that disorders higher in this order should be diagnosed over disorders lower in the hierarchy. Thus, in terms of standard diagnostic practice, the presence of organic mental disorders should be ruled out first, then the schizophrenic disorders, then affective disorders, etc. This principle of diagnostic precedence is analogous to the authority relationship among different levels of rank in the military.

Notice that the pecking order relationship among these five general mental disorders also carries another implication. If a patient has an organic mental disorder, the patient can (and often does) have the symptoms associated with disorders that are lower in the hierarchy. Thus, a patient with Alzheimer's disease can develop hallucinations like a schizophrenic, can have marked sleep disturbance like someone who is depressed, can develop a fear of leaving the house like someone with anxiety disorder, and show the rigidity and need to be controlled like someone with an obsessive-compulsive personality disorder. However, a patient with anxiety disorder such as agoraphobia should not show the disturbed memory patterns and disorientation of a patient with an organic mental disorder.

The important point to note is that disorders placed higher in this pecking order view of hierarchy can explain any of the symptoms of disorders lower in the hierarchy; however, the reverse should not occur. There should be symptoms that will be manifest in patients with schizophrenia that will not occur in patients with personality disorders.

The pecking order approach to the hierarchical arrangement of mental disorder categories was popularized by Foulds and Bedford (1975). The specifics of their approach to the classification of mental disorders differs from that presented above, but the general outline is the same. An important corollary to this pecking order view of the hierarchical arrangement among mental disorder categories is that this view suggests that there will be a strong severity dimension in any descriptive approach to the classification of these disorders. Mental disorders higher in this system will be associated with many more symptoms than are mental disorders lower in this system. Descriptive studies of psychopathology have repeatedly found a strong severity dimension that the pecking order view would predict.

4.03.9.4 Problem of Diagnostic Overlap/ Comorbidity

When discussing the categorical model of classification, one of the tenets that was attributed to the model stated: "In the borderline areas where categories may overlap, the number of overlapping patients should be relatively small." Diagnostic overlap refers to the relative percentage of patients with one diagnosis who also meet the criteria for another diagnosis. As the tenet above states, some diagnostic overlap is expected. But the relative amount of overlap should be small.

One terminological note should be made before proceeding. The literature on this issue is grouped under the general heading of comorbidity. This term is from the medical literature, because it is well recognized that some medical disorders tend to go together. For instance, individuals who develop AIDS are relatively likely to develop yeast infections, sarcomas, and other disorders because of their compromised autoimmune system. The term comorbidity refers to the pattern of co-occurrences of these medical disorders. However, because the concept of comorbidity implies the acceptance of a disease model, the preferred term in this chapter will be "diagnostic overlap."

One of the earliest studies that focused on diagnostic overlap was by Barlow, DiNardo, Vermilyea, Vermilyea, and Blanchard (1986).

These investigators reported on 126 patients who were referred for the treatment of anxiety. These patients were administered structured interviews. Of the 126 patients interviewed, 108 were assigned one of seven diagnoses that fit within the anxiety/affective disorder spectrum. Of these 108 patients 65% were given at least one additional diagnosis. This is a high level of diagnostic overlap and apparently was much higher than these researchers had expected *a priori*.

A large number of other empirical studies have confirmed the high levels of diagnostic overlap using the *DSM-III* and subsequent classifications. Many of these studies are discussed in an excellent review by Clark, Watson, and Reynolds (1995). Examples of the results found in their review are listed below. For example, these reviewers noted one study of personality disorder diagnoses in a state hospital population that found that these patients met the diagnostic criteria for an average of 3.75 Axis II disorders. In addition to the personality disorders, the depressive disorders also have striking overlap with many other disorders. For instance, over half of the patients with major depressive disorder as well as patients with dysthymic disorder were found to have at least one co-occurring mental disorder. Depression shows significant overlap even with disorders that one might not expect overlap. For instance, a sample of antisocial personality disorder patients showed that one-third of these individuals also had a depressive diagnosis (Dinwiddie & Reich, 1993). These antisocial patients, less surprisingly, also had high rates of alcoholism (76%) and other substance use (63%). Even broad epidemiological studies on normal community samples show high rates of diagnostic overlap. Clark et al. reported that, in two national studies, over half of the individuals who had one mental disorder diagnosis had at least one more diagnosis. In some samples, the rate of diagnostic overlap is even higher. For instance, a study of suicidal patients showed that these individuals averaged about four mental disorders (Rudd, Dahm, & Rajab, 1993).

Together, these and many other studies suggest that the number of overlapping patients among mental disorder diagnoses is not small at all. Instead, diagnostic overlap is a standardly occurring phenomenon. Blashfield, McElroy, Pfohl, and Blum (1994) studied 151 patients who had been administered a structured interview to assess personality disorders. In this sample, only 24% of the patients met the criteria for one and only one personality disorder. Exactly the same percentage met the diagnostic criteria for at least four personality disorders,

with 6% meeting the criteria for six of the eleven *DSM-III-R* personality disorders! When Blashfield et al. attempted to identify prototypic patients (i.e., individuals who met at least eight of the diagnostic criteria for a specific disorder), they found that only 15% of the patients would qualify as prototypic. However, most of these individuals also satisfied the diagnostic criteria for other disorders. When a prototypic patient was defined as an individual with eight or more criteria for a personality disorder and the lack of an additional personality diagnosis, only 1% of the patients were prototypes. In effect, patients with mixed symptom patterns are much more typical than are patients who represent relatively pure forms of a disorder.

Clark, Watson and Reynolds (1995) suggested that there are three related issues associated with the problem of diagnostic overlap. The first issue concerns the hierarchical organization of categories. As discussed earlier, one view of hierarchy is a pecking order approach. This view was implicitly adopted by the *DSM-III* because a number of exclusion rules were included in the diagnostic criteria for different disorders, so that a lower diagnosis would not be made if a higher order diagnosis was possible. However, research on the exclusionary criteria suggested that these criteria were arbitrary. Exclusionary criteria were mostly deleted in the *DSM-III-R* and the *DSM-IV*.

The second issue associated with the diagnostic overlap issue concerns the heterogeneity with diagnostic categories. In effect, the extensive diagnostic overlap suggests that the definitions of the various mental disorders are too broad and inclusive. Evidence of excessive heterogeneity comes from other sources, according to Clark et al. For instance, direct studies of the variability in symptom patterns have shown high levels of variability within disorders such as schizophrenia, depression, eating disorders and anxiety disorders. Another line of evidence of heterogeneity is the frequency with which mixed or atypical diagnoses are used. For instance, Mezzich, Fabrega, Coffman, and Haley (1989) found that the majority of patients with a dissociative disorder fit the criteria for an atypical dissociative disorder.

The third issue associated with the comorbidity finding is the increasing support associated with replacing the categorical approach to classification with a dimensional view. In the discussion of these two models, it was noted that the dimensional model is the simpler of the two. Unless there is clear evidence of the existence of discrete categories, a dimensional approach is the more parsimonious.

A number of researchers, when confronted with high rates of diagnostic overlap, have suggested that a dimensional model should be used. For instance, in the area of the personality disorders where some of the highest rates of diagnostic overlap have been found, interest in a "Big Five" dimensional approach to the personality disorders has been attracting increasing support. Another example in which a dimensional model is gaining popularity is in the subclassification of schizophrenia. The classic Kraepelinian subtypes do not generate sufficiently homogeneous categories. Various dimensional schemes (process-reactive, paranoid-nonparanoid, positive vs. negative symptoms) have more empirical support than the use of categorical subtypes.

4.03.10 CONCLUDING COMMENTS

This chapter provides a rather simplified overview of the issues associated with the classification of psychopathology. It attempts to help the reader gain a better understanding of classification by starting with a reasonably detailed history of classificatory systems, and thereby give some idea how many of the features of contemporary classificatory systems have evolved. The text also presents a succinct and readable presentation of four issues that currently attract a reasonable degree of attention in the literature. However, justice has not been done to many of the other complex issues that face both clinicians and scientists interested in classification, such as whether to use semistructured interviews as the "gold standard" for measurement (Mezzich, 1984), the role of values in the diagnostic practice of mental health professionals, whether or not certain mental disorders are sexually or racially biased (Busfield, 1996; Nuckolls, 1992; Widom, 1984), the relevance of life-span measures of psychopathology (Roff & Ricks, 1970), and the problem of focusing on the individual patient as the basic unit of psychopathology (as opposed to families or systems or interpersonal relationship patterns) (Clarkin & Miklowitz, 1997; Francis, Clarkin & Ross, 1997; Williams, 1997).

Like any general topic, the classification of psychopathology becomes a very complex topic when analyzed in detail.

Perhaps I believe that the world can get forward most by clearer and clearer definitions of fundamentals. Accordingly I propose to stick to the tasks of nomenclature and terminology, unpopular and ridicule-provoking though they may be. (Southard, as quoted by Menninger, 1963, p. 3)

4.03.11 REFERENCES

- American Psychiatric Association (1933). Notes and comments: Revised classified nomenclature of mental disorders. *American Journal of Psychiatry*, 90, 1369-1376.
- American Psychiatric Association (1952). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: Author.
- American Psychiatric Association (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.
- American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. Rev.). Washington, DC: American Psychiatric Press.
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Press.
- Alexander, F. G., & Selesnick, S. T. (1966). *The history of psychiatry*. New York: Harper and Row.
- Austin, T. J. (1859/1976). *A practical account of general paralysis*. New York: Arno Press.
- Barlow, D. H., DiNardo, P. A., Vermilyea, B. B., Vermilyea, J., & Blanchard, E. B. (1986). Co-morbidity and depression among the anxiety disorders: Issues in diagnosis and classification. *Journal of Nervous and Mental Disease*, 174, 63-72.
- Barsalou, L. W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Hillsdale, NJ: Erlbaum.
- Bayer, R. (1981). *Homosexuality and American psychiatry*. New York: Basic Books.
- Berrios, G. E. (1993). Personality disorders: A conceptual history. In P. Tyrer & G. Stein (Eds.), *Personality disorder reviewed* (pp. 17-41). London: Gaskell.
- Berrios, G. E., & Hauser, R. (1988). The early development of Kraepelin's ideas on classification: A conceptual history. *Psychological Medicine*, 18, 813-821.
- Blashfield, R. K. (1991). Models of psychiatric classification. In M. Hersen & S. M. Turner (Eds.), *Adult psychopathology and diagnosis* (2nd ed. pp. 3-22). New York: Wiley.
- Blashfield, R. K., McElroy, R. A., Pfohl, B., & Blum, N. (1994). Comorbidity and the prototype model. *Clinical Psychology: Science and Practice*, 1, 96-99.
- Blustein, B. E. (1991). *Preserve your love for science: Life of William A. Hammond, American neurologist*. Cambridge, UK: Cambridge University Press.
- Braslow, J. T. (1995). Effect of therapeutic innovation on perception of disease and the doctor-patient relationship: A history of general paralysis of the insane and malaria fever therapy, 1910-1950. *American Journal of Psychiatry*, 152, 660-665.
- Busfield, J. (1996). *Men, women and madness: Understanding gender and mental disorder*. New York: New York University Press.
- Campbell, E. J. M., Scadding, J. G., & Roberts, R. J. (1979). The concept of disease. *British Medical Journal*, 2, 757-762.
- Cantor, N., Smith, E. E., French, R. D., & Mezzich, J. (1980). Psychiatric diagnosis as a prototype categorization. *Journal of Abnormal Psychology*, 89, 181-193.
- Clark, L. A., Watson, D., & Reynolds, S. (1995). Diagnosis and classification of psychopathology: Challenges to the current system and future directions. *Annual Review of Psychology*, 46, 121-153.
- Clarkin, J. F., & Mikowitz, D. J. (1997). Marital and family communication difficulties. In T. A. Widiger, A. J. Francis, H. A. Pincus, R. Ross, M. B. First, & W. Davis (Eds.), *DSM-IV sourcebook* (Vol. 3, pp. 631-672). Washington, DC: American Psychiatric Press.

- Dinwiddie, S. H., & Reich, T. (1993). Attribution of antisocial symptoms in coexistent antisocial personality disorder and substance abuse. *Comprehensive Psychiatry*, 34, 235–242.
- Essen-Moller, E. (1971). Suggestions for further improvement of the international classification of mental disorders. *Psychological Medicine*, 1, 308–311.
- Everitt, B. S. (1974). *Cluster analysis*. New York: Halstead Press.
- Feighner, J. P., Robins, E., Guze, S., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, 143, 57–63.
- Feinstein, A. R. (1967). *Clinical judgment*. Huntington, VA: Krieger.
- Foulds, G. A., & Bedford A. (1975). Hierarchy of classes of personal illness. *Psychological Medicine*, 5, 181–192.
- Francis, A. J., Clarkin, J. F., & Ross, R. (1997). Family/relational problems. In T. A. Widiger, A. J. Francis, H. A. Pincus, R. Ross, M. B. First, & W. Davis (Eds.), *DSM-IV sourcebook* (Vol. 3, pp. 521–530). Washington, DC: American Psychiatric Press.
- Goffman, E. (1961). *Asylums*. London: Penguin.
- Grinker, R. R., Werble, B., & Drye, R. C. (1968). *The borderline syndrome*. New York: Basic Books.
- Hammond, W. A. (1883). *A treatise on insanity in its medical relations*. New York: Appleton.
- Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: Free Press.
- Hull, D. L. (1988). *Science as a process*. Chicago: University of Chicago Press.
- Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. New York: Wiley.
- Kendell, R. E. (1975). *The role of diagnosis in psychiatry*. Oxford, UK: Blackwell.
- Kendell, R. E. (1976). The concept of disease. *British Journal of Psychiatry*, 128, 508–509.
- Kendell, R. E. (1979). Alcoholism: A medical or a political problem. *British Medical Journal*, 1, 367–381.
- Kendell, R. E. (1986). What are mental disorders? In A. M. Freedman, R. Brotman, I. Silverman, & D. Huston (Eds.), *Issues in psychiatric classification* (pp. 23–45). New York: Human Sciences Press.
- Kendell, R. E. (1991). Relationship between the DSM-IV and ICD-10. *Journal of Abnormal Psychology*, 100, 297–301.
- Kirk, S. A., & Kutchins, H. (1992). *The selling of DSM: The rhetoric of science in psychiatry*. Hawthorne, NY: Walter deGruyter.
- Klerman, G. L. (1978). The evolution of a scientific nosology. In J. C. Shershow (Ed.), *Schizophrenia: Science and practice* (pp. 99–121). Cambridge, MA: Harvard University Press.
- Kraepelin, E. (1902/1896). *Clinical psychiatry: A text-book for students and physicians* (6th ed., translated by A. R. Diefendorf). London: Macmillan.
- Kreitman, N. (1961). The reliability of psychiatric diagnosis. *Journal of Mental Science*, 107, 878–886.
- Lacan, J. (1977). *Ecrits: A selection*. New York: Norton.
- Laing, R. D. (1967). *The politics of experience*. London: Penguin.
- Lorr, M. (1966). *Explorations in typing psychotics*. New York: Pergamon.
- Matza, D. (1969). *Becoming deviant*. Englewood Cliffs, NJ: Prentice-Hall.
- Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, 50, 266–275.
- Menninger, K. (1963). *The vital balance*. New York: Viking.
- Mezzich, J. E. (1984). Diagnosis and classification. In S. M. Turner & M. Hersen (Eds.), *Adult psychopathology and diagnosis* (pp. 3–36). New York: Wiley.
- Mezzich, J. E., Fabrega, H., Coffman, G. A., & Haley, R. (1989). DSM-III disorders in a large sample of psychiatric patients: Frequency and specificity of diagnosis. *American Journal of Psychiatry*, 146, 212–219.
- Mezzich, J. E., Fabrega, H., Mezzich, A. C., & Coffman, G. A. (1985). International experience with DSM-III. *Journal of Nervous and Mental Disease*, 173, 738–741.
- Nelson, G., & Platnick, N. (1981). *Systematics and biogeography: Cladistics and vicariance*. New York: Columbia University Press.
- Nuckolls, C. W. (1992). Toward a cultural history of the personality disorders. *Social Science and Medicine*, 35, 37–47.
- Plaut, F. (1911). *The Wasserman sero-diagnosis of syphilis in its application to psychiatry*. New York: Journal of Nervous and Mental Disease Publishing Company.
- Project Match Research Group (1997). Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes. *Journal of Studies on Alcohol*, 58, 7–29.
- Quetel, C. (1990). *History of syphilis*. Baltimore: Johns Hopkins University Press.
- Raines, G. N. (1952). Forward. In American Psychiatric Association, *Diagnostic and statistical manual of mental disorders* (1st ed., pp. v–xi). Washington, DC: American Psychiatric Association.
- Regier, D. A., Kaelber, C. T., Roper, M. T., Rae, D. S., & Sartorius, N. (1994). The ICD-10 clinical field trial for mental and behavioral disorders: Results in Canada and the United States. *American Journal of Psychiatry*, 151, 1340–1350.
- Reznek, L. (1987). *The nature of disease*. London: Routledge & Kegan Paul.
- Roff, M., & Ricks, D. F. (Eds.) (1970). *Life history research in psychopathology*. Minneapolis, MN: University of Minnesota Press.
- Rosenhan, D. L. (1973). On being sane in insane places. *Science*, 179, 250–258.
- Roth, A., & Fonagy, P. (1996). *What works for whom: A critical review of psychotherapy research*. New York: Guilford.
- Rudd, M. D., Dahm, P. F., & Rajab, M. H. (1993). Diagnostic comorbidity in persons with suicidal ideation and behavior. *American Journal of Psychiatry*, 147, 1025–1028.
- Russell, J. A. (1991). In defense of a prototype approach to emotion concepts. *Journal of Personality and Social Psychology*, 60, 37–47.
- Russell, J. A., & Fehr, B. (1994). Fuzzy concepts in a fuzzy hierarchy: Varieties of anger. *Journal of Personality and Social Psychology*, 67, 186–205.
- Rutter, M., Shaffer, D., & Shepard, M. (1975). *A multiaxial classification of child psychiatry disorders*. Geneva, Switzerland: World Health Organization.
- Sadler, J. Z., Wiggins, O. P., & Schwartz, M. A. (1994). *Philosophical perspectives on psychiatric diagnostic classification*. Baltimore: Johns Hopkins University Press.
- Sarbin, T. R. (1997). On the futility of psychiatric diagnostic manuals (DSMs) and the return of personal agency. *Applied and Preventive Psychology*, 6, 233–243.
- Scadding, J. G. (1959). Principles of definition in medicine with special reference to chronic bronchitis and emphysema. *Lancet*, 1, 323–325.
- Scadding, J. G. (1969). Diagnosis: The clinician and the computer. *Lancet*, 2, 877–882.
- Skinner, H. A. (1986). Construct validation approach to psychiatric classification. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 307–331). New York: Guilford Press.
- Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco: Freeman.
- Spitzer, R. L., Endicott, J., & Robins, E. (1975). Research

- diagnostic criteria. *Archives of General Psychiatry*, 35, 773–782.
- Spitzer, R. L., & Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, 125, 341–347.
- Spitzer, R. L., & Williams, J. B. W. (1987). Revising DSM-III: The process and major issues. In G. L. Tischler (Ed.), *Diagnosis and classification in psychiatry* (pp. 425–434). New York: Cambridge University Press.
- Spitzka, E. C. (1883). *Insanity: Its classification, diagnosis and treatment*. New York: Bermingham.
- Stangl, D., Pfohl, B., Zimmerman, M., Bowers, W., & Corenthal, C. (1985). Structured interview for the DSM-III personality disorders. *Archives of General Psychiatry*, 42, 519–596.
- Stengel, E. (1959). Classification of mental disorders. *Bulletin of the World Health Organization*, 21, 601–663.
- Stromgren, E. (1991). A European perspective on the conceptual approaches to psychopathology. In A. Kerr & H. McClelland (Eds.), *Concepts of mental disorders: A continuing debate* (pp. 84–90). London: Gaskell.
- Szasz, T. (1961). *The myth of mental illness*. New York: Hoeber-Harper.
- Veith, I. (1965). *Hysteria: The history of a disease*. Chicago: University of Chicago Press.
- Wakefield, J. C. (1992). The concept of mental disorder: On the boundary between biological facts and social values. *American Psychologist*, 47, 373–388.
- Wakefield, J. C. (1993). Limits of operationalization: A critique of Spitzer and Endicott's (1978) proposed operational criteria for mental disorder. *Journal of Abnormal Psychology*, 102, 160–172.
- Wakefield, J. C. (1997a). Diagnosing DSM-IV—Part I: DSM-IV and the concept of disorder. *Behavioral Research and Therapy*, 35, 633–649.
- Wakefield, J. C. (1997b). Diagnosing DSM-IV—Part II: Eysenck (1986) and the essentialist fallacy. *Behavioral Research and Therapy*, 35, 651–665.
- Walker, L. (1987). Inadequacies of the masochistic personality disorder diagnosis for women. *Journal of Personality Disorders*, 1, 183–189.
- Widom, C. (Ed.) (1984). *Sex, roles and psychopathology*. New York: Plenum.
- Williams, J. B. W. (1997). The DSM-IV multiaxial system. In T. A. Widiger, A. J. Francis, H. A. Pincus, R. Ross, M. B. First, & W. Davis (Eds.), *DSM-IV sourcebook* (Vol. 3). Washington, DC: American Psychiatric Press.
- Wing, J. K. (1978). *Reasoning about madness*. Oxford, UK: Oxford University Press.
- Woodruff, R. A., Goodwin, D. W., & Guze, S. B. (1974). *Psychiatric diagnosis*. New York: Oxford University Press.
- World Health Organization (1948). *Manual of the international statistical classification of diseases, injuries, and causes of death*. Geneva, Switzerland: Author.
- World Health Organization (1957). Introduction to *Manual of the international statistical classification of diseases, injuries, and causes of death* (7th ed.). Geneva, Switzerland: Author.
- World Health Organization (1978). *Mental disorders: Glossary and guide to their classification in accordance with the ninth revision to the International Classification of Diseases*. Geneva, Switzerland: Author.
- World Health Organization (1992). *The ICD-10 classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines*. Geneva, Switzerland: Author.
- World Health Organization (1993). *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. Geneva, Switzerland: Author.
- Wulff, H. R., Pedersen, S. A., & Rosenberg, R. (1986). *Philosophy of medicine*. Boston: Blackwell Scientific.
- Wyer, R. S., & Srull, T. K. (1989). *Memory and cognition in a social context*. Hillsdale, NJ: Erlbaum.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.
- Zubin, J. (1967). Classification of behavior disorders. *Annual Review of Psychology*, 28, 373–406.

4.04

Clinical Interviewing

EDWARD L. COYLE

Oklahoma State Department of Health, Oklahoma City, OK, USA
and

DIANE J. WILLIS, WILLIAM R. LEBER, and JAN L. CULBERTSON
University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

4.04.1 PURPOSE OF THE CLINICAL INTERVIEW	82
4.04.1.1 <i>Gathering Information for Assessment and Treatment</i>	82
4.04.1.2 <i>Establishing Rapport for Assessment and Treatment</i>	82
4.04.1.3 <i>Interpersonal Style/Skills of the Interviewer</i>	82
4.04.1.4 <i>Structuring the Interview</i>	83
4.04.1.4.1 <i>Setting variables</i>	83
4.04.1.4.2 <i>Preparing for the patient</i>	84
4.04.1.5 <i>Introductory Remarks</i>	84
4.04.1.6 <i>How to Open the Interview</i>	85
4.04.1.7 <i>The Central Portion of the Interview</i>	85
4.04.1.8 <i>Closing the Interview</i>	86
4.04.1.9 <i>The Collateral Interview</i>	86
4.04.2 DEVELOPMENTAL CONSIDERATIONS IN INTERVIEWING	87
4.04.2.1 <i>Interviewing Children (Preschool Age through Older Elementary)</i>	87
4.04.2.2 <i>Interviewing Parents</i>	87
4.04.2.3 <i>Social Context</i>	88
4.04.2.4 <i>Developmental Context</i>	89
4.04.2.5 <i>Direct Interview of Children</i>	89
4.04.2.6 <i>Adolescents</i>	90
4.04.2.6.1 <i>Separation-individuation</i>	91
4.04.2.6.2 <i>Resolving conflict with authority figures</i>	91
4.04.2.6.3 <i>Peer group identification</i>	91
4.04.2.6.4 <i>Realistic appraisal and evaluation of self-qualities</i>	92
4.04.2.7 <i>Interviewing Young Adults (18–40 Years)</i>	92
4.04.2.8 <i>Interviewing Adults in Middle Adulthood (40–60 Years)</i>	93
4.04.2.9 <i>Interviewing Older Adults (60–70 Years)</i>	93
4.04.2.10 <i>Interviewing In Late Adulthood (70 Years–End of Life)</i>	93
4.04.3 INTERVIEWING SPECIAL POPULATIONS OR SPECIFIC DISORDERS	94
4.04.3.1 <i>Interviewing Depressed Patients</i>	94
4.04.3.2 <i>Interviewing Anxious Patients</i>	94
4.04.4 SUMMARY	95
4.04.5 REFERENCES	96

4.04.1 PURPOSE OF THE CLINICAL INTERVIEW

The clinical interview is extremely important as a diagnostic tool in the assessment and treatment of patients. Clinicians who do thorough and competent interviews have a much better understanding of the developmental course of symptoms presented by the patient. Indeed, before there were any personality inventories, before the Rorschach and one-way mirror behavioral observation, there was the clinical interview. The purpose of the clinical interview is to gain sufficient information from the informant or informants to formulate a diagnosis, assess the individual's strengths and liabilities, assess the developmental and contextual factors that influence the presenting concerns, and to allow planning for any interventions to follow. The interview is in many instances the ultimate clinical tool, and effective interviewing must be an integral part of any clinician's professional abilities. Although the clinical interview is used primarily to gather information for clinical evaluation or psychotherapeutic treatment, it can also serve the purpose of preparing the patient for therapy, and less frequently the interview process itself provides some relief from psychological distress. The interview may be performed in many different settings: outpatient private practice, community mental health center, psychiatric hospital, prison, emergency or medical hospital room, school, and others.

While the amount of time devoted to the interview, the setting, and the purposes may vary, the features of an effective clinical interview remain the same. When completed, the interviewer has created a relatively comprehensive portrait of the patient that can be communicated readily to others and will provide the basis for making important judgments about the subject of the interview. The relative importance of various symptoms or concerns should be established, along with an estimate of the individual's overall functioning. The relative importance of various symptoms or concerns should be established, and some estimate of the individual's responses in a variety of settings can be made with an acceptable degree of validity. These features can be said to be a part of any clinical interview. Some specific purposes of the interview are described next, along with suggestions about different approaches and emphases the clinician may be required to take.

4.04.1.1 Gathering Information for Assessment and Treatment

The most common purposes of the clinical interview are to gather information to establish a

diagnosis, evaluate mental status and historical data that impact upon the individual, and provide a full understanding of the important personality, biological, and environmental variables that have brought the patient to this point. All treatment planning begins with some type of formal or informal evaluation. The clinical interview is the most effective way to gain an understanding of the current functioning and the difficulties faced by the patient, and is a necessary adjunct to the data gathered from other assessment approaches. In the clinical interview, the clinician inquires directly and in a focused manner about the patient's development, adaptation, and current difficulties. When the interview is part of a comprehensive evaluation, some features may be emphasized as a result of the specific referral reason that would not be as prominent in the interview conducted for psychotherapy treatment planning. As an example, the interview conducted for an initial psychoeducational evaluation of an elementary grade school child to determine reasons for school failure is likely to entail considerable emphasis upon academic and learning history and the involvement of one or more of the child's teachers as collateral informants. If the same child were to present later for psychotherapeutic interventions to address the depression and oppositional behaviour problems identified by previous evaluation to be the cause of his academic failure, the interviewer would likely spend more time and effort in determining family interactions, parenting skills, and social supports.

4.04.1.2 Establishing Rapport for Assessment and Treatment

One function of the clinical interview is to prepare the patient for the clinical interventions that follow, including additional formal assessment procedures. In order to obtain valid psychometric data, the patient must be adequately cooperative and invested in the testing process. The interview can help the clinician achieve this end by providing a sense of professional intimacy and a feeling of compassion and interest in the patient's well-being. Thus prepared, the respondent is more willing to give themselves over to the process, and to perceive it as being something that will provide them with some beneficial outcome.

4.04.1.3 Interpersonal Style/Skills of the Interviewer

While the basic purpose of gathering relatively concrete information may be accomplished by individuals with a minimum of

training and sensitivity, there are a number of personal qualities that tend to improve the quality of information gained and to result in a more helpful and pleasant experience on the part of the informant. Chief among these is the quality of empathy, which must be readily recognized by the informant through subtle and overt communications from the interviewer.

Empathy means identifying with and understanding someone else's feelings, motives, or word view. It means entering the private perceptual world of another person and being at home in it—to be sensitive to what they feel (Egan, 1994; Luborsky, 1996). An intellectual understanding of empathy, however, does not provide one with the interpersonal skills and experience that result in the ability to truly resonate to the informant's experience and to respond in ways that will ease the flow of information of a personal and often sensitive nature. The skill and art of attuning oneself not only to the overt communications of the patient, but also to the underlying feelings and meanings, must become a continuing focus of attention for the interviewing clinician. While much of this process is not fully accessible to conscious awareness, there are some components that lend themselves readily to examination and modification. For example, the interviewer's responses that communicate to the informant negative value judgments are perhaps more easily modified.

Although the mental health fields and their practitioners have often been vilified for their purported moral relativism, no reasonable clinician would believe himself or herself to be free of individual prejudices and deeply-held convictions regarding right and wrong. These values are a part of each person, and to truly expunge them would result in an insipid and ineffective shell of a human being. The relevance to a discussion of clinical interviewing is this: the effective interviewer takes care to be aware of his or her own expectations and biases regarding human behaviour and strives to avoid making explicit negative judgments of the informant in order to provide a comfortable and supportive environment during the interview. This skill can be and is developed and improved through careful attention to the process, internal changes within the interviewer during the interaction, and by effective supervision and review of actual interviews with other clinicians.

Often such judgments can be communicated to the respondent with no more than a change in facial expression or in a shift in questioning. Specific wording and follow-up questions sometimes can have the effect of casting a chill upon the interview process. For example, the interviewer who learns the informant is homo-

sexual and then avoids asking questions about sexual functioning and current relationships readily communicates their discomfort with an important aspect of the informant's personality. The effective interviewer does not perfect an unexpressive mask, but does develop the ability to decrease the immediate translation of visceral responses into explicit behaviours during the interview. Introspection about areas that increase the clinician's anxiety and honest confrontation about discriminatory beliefs are necessary if one is to perform clinical tasks competently and ethically.

4.04.1.4 Structuring the Interview

As with any therapeutic or evaluative intervention, the setting and structure of the interview have a significant effect on the outcome of the interaction. Because the actual face-to-face time spent with patients must be as productive and positive as possible, the clinician should take care to prepare for the clinical interview prior to contact. While the goals of the interview may vary somewhat as discussed previously, many factors common to all clinical interviews should at least be in one's mind prior to and during the interview.

4.04.1.4.1 Setting variables

Some basic attention should be given to simple environmental factors when preparing for the interview. Although many fruitful interviews have been conducted with patients, families, and other sources under conditions that might be charitably described as less-than-optimal, doing so in a comfortable and soothing environment will often add to an informant's ease in discussing delicate and/or emotionally charged matters. Seating accommodations should be given consideration, as hard, uncomfortable, or rickety, precarious seats can add a tinge of anxiety or discomfort to the gestalt of the interview process and thus to the evaluation and/or treatment that follows the interview. It should go without saying that the space being used for the clinical interview should be held relatively inviolate from intrusions, including external noises and conversation to the degree possible. While most people are able to tolerate minor interruptions such as a ringing telephone, having another clinician open the door while your patient is tearfully recounting a past trauma is likely to be somewhat harmful to the tentative alliance you are developing. Therefore, if you work in a setting with multiple users you will do well to take precautions to avoid such disruptions. A white-sound generator can help

decrease the penetration of external sounds and add somewhat to the intimacy of the interaction. Throughout the interview the clinician will be carefully observing the behaviours of the subject, noting congruencies and incongruencies, attending to shifts in voice and posture. One sometimes overlooked source of information that may add to the interview process is that of behavioral observations made while the patient and collaterals are in the waiting area of the clinic or office. Often it is possible to observe interactions and general demeanor while you organize paperwork or make other preparations before formally introducing yourself. It may then be helpful to comment during interview on some salient interaction or response. Of course, as has been particularly noted by various custody and forensic evaluators (Bricklin, 1990), the behavior in the waiting room must not be taken as necessarily representative of the person's usual response outside of the clinic. However, one often can observe telling interactional patterns, particularly between parents and children, and this may provide opportunity for addressing problematic areas during the subsequent interview.

4.04.1.4.2 Preparing for the patient

It is common practice now to present the patient with relatively extensive questionnaires or personal history forms prior to the first clinic visit. With this information in hand, the clinician may be able to focus in quickly on the salient symptomatology and current concerns. When available, this information should be used to tailor the interview, allotting time during the face-to-face interview in the most effective manner. If the clinician does choose to utilize such instruments, he or she would be well served to take the time necessary to review the data prior to entering the room with the informant. Watching the professional read and murmur over the forms for minutes while the informant is ignored or asked disconnected questions can be expected to result in a sense of devaluation for the informant. It also gives the impression of disorganization and lack of preparation on the part of the clinician. Neither of these will be helpful in the ensuing interview process.

4.04.1.5 Introductory Remarks

It is helpful to develop a standard approach to the clinical interview, including the introduction and beginning of the interview. One should introduce oneself, and give the informant information about the purpose of the

interview and the expected duration. The interviewer's role and title should be clarified, and any supervisory or other training relationship must be disclosed prior to beginning the interview. It is essential that issues of confidentiality be fully addressed, and the informant be given opportunity and encouragement to ask questions about disclosure of information. If any of the data obtained will be shared with other individuals, this must be explained clearly. This is of particular importance in forensic or custody evaluations. When interviewing children and parents, keep in mind the fact that in many jurisdictions the noncustodial parent may retain full rights to examine medical records, including data from the clinical interview. Even if the informant has signed a general disclosure or consent for treatment, it is the clinician's ethical responsibility to review duties to warn and the possible limits on confidentiality. The legal definition of informed consent in many jurisdictions is not necessarily satisfied by the presence of a signature on a form, but rather is established by questioning the informant about their understanding at the time the information was given. The best practice is for the clinician to do their best to make certain that the person with whom they are communicating for professional purposes is fully informed of such issues. In illustration, imagine for a moment being 30 minutes into an interview with a man who informs you very clearly that he intends to use the pistol in his car to shoot his wife when he returns home. If you have fully informed him of the limits of confidentiality, you are in a very distressing situation. If you have not done so, your position is much worse. The growth of managed care and its attendant prospective treatment review process may complicate the ethical duties involved in the clinical interview. As Corcoran and Vandiver (1996) point out, "There can be no doubt that managed care has restricted clients autonomy and interferes with the confidential relationship" (p. 198). During the initial interview and prospective utilization review of a patient whose care is managed by a third (or possibly fourth) party, the clinician may find him or herself in the uncomfortable position of being more the agent of the managed-care organization than the advocate of the patient. In such relationships, it is imperative that the patient be made fully aware at the outset of the interview of the additional limits of confidentiality imposed by the managed-care entity. This may include multiple reviews of the data gained during interview and any subsequent treatment sessions. An additional ethical concern arises in the clinical interview with regard to the establishment of a professional relationship

and responsibility for the clinical care of the patient. Does performing the clinical interview and prospective review obligate the clinician to provide service even if the managed-care entity denies authorization? Again, the only way to avoid difficulties, misunderstandings, and possible litigation or board complaints is to be absolutely clear with interviewees and any involved third-party payer about these issues prior to the professional contact. If it is possible that the interviewing clinician will not receive reimbursement from the managed-care company for services, any alternative financial arrangements should also be discussed with the prospective patient before any formal clinical contact. If there are inherent limitations to the number of sessions or type of interventions that are covered by the third-party payer, the potential client should also be made aware of these before ending the interview. Of course, it is possible that no treatment will be necessary; thus it seems sensible to leave discussing the mechanics of paying for it until it is determined to be needed.

4.04.1.6 How to Open the Interview

The best way to open the interview is with a very general, open-ended question about the circumstances that have brought the patient to the interview. Morrison (1995) recommends taking approximately eight to 10 minutes in the typical one-hour interview to allow the respondent to explain in their own words their needs and history. Morrison points out that, among other things, this provides the clinician an opportunity to obtain a true flavor for the respondent's personality and communication style, and to make general observations of behavior, affect, and thought process relatively free from the clinician's direction. An example of an opening question might be "Please tell me about the things that are concerning you most right now" or "I would like for you to tell me what you need some assistance with now" or even "Please give me an idea of how you came to be here today." The amount of information gathered during this portion of the interview will be to some degree dependent upon the respondent's intellectual ability and verbal facility. Many people are characterologically unwilling to self-disclose, even within the confines of the clinical interview, and may require additional urging. The clinician should generally respond to hesitations with supportive restatement of the opening question, or with gentle encouragement and reflection of any apprehension that is detected. If hesitation or lack of content appear to be due to cognitive

limitations, disorientation, or distractibility, it may be helpful to ask more direct and close-ended questions based upon previously obtained information or the patient's brief verbalizations. It is generally not desirable to lead the patient any more than necessary, as the more you query the less likely you will be able to distinguish between accurate responses and those that are colored by the demands experienced by the patient. However, in some cases the clinician must take a more directive approach to complete the interview successfully.

The topics to be included in every interview are:

- (i) Introduction, purpose of interview, credentials/role of interviewer;
- (ii) Confidentiality and exceptions;
- (iii) Presenting problems (preferably phrased in general, open-ended manner);
- (iv) Mood/Anxiety symptoms;
- (v) Impulse control and direct inquiry of suicidal ideation/history;
- (vi) Current social, academic, and vocational functioning;
- (vii) System of social support;
- (viii) Environmental factors, including current basic needs/shelter;
- (ix) Developmental factors (especially for children) that may influence symptom presentation;
- (x) Medical history, including family health history and previous treatment/hospitalization;
- (xi) Substance use;
- (xii) Legal involvement and history; and
- (xiii) Vegetative symptoms.

4.04.1.7 The Central Portion of the Interview

After the initial introduction, housekeeping, and rapport-building, it is time to focus upon the most salient features of the person being evaluated, and the circumstances that maintain the current dysfunction. Once the presenting problems have been identified and an adequate alliance with the respondent established, the clinician must utilize their knowledge of psychopathology and diagnostic criteria to fully understand and classify the presenting problems, as well as to identify the primary strengths and resources that will be drawn upon by the patient and the professionals involved in subsequent interventions. The central portion of the interview is dedicated to adding to the framework established by queries about the presenting problem. One mistake made by novice (as well as by some more seasoned but overly concrete) interviewers is to rigidly adhere to an interviewing framework, disregarding the natural flow of conversation.

If one is unable to recognize the more subtle verbal and nonverbal messages that should be probed and instead forces one's way forward, the clinician will end up with less information than they should. Thus, it is essential to attend carefully to shifts in mood during the interview, both within the patient and the interviewer.

Luborsky (1996) details the utilization of momentary shifts in mood during a therapy session to focus upon vital underlying thoughts that are salient to the therapeutic issues. The interviewing clinician can also benefit by noticing changes in voice tone, volume, and content of speech. During the central portion of the interview the clinician continues to focus on the problems and possible explanations for present distress. When possible, avoid becoming involved in digressive topics, as some respondents may prefer to spend most of the available time presenting problems that are not central to the services being sought. By the same token, it is the clinician's responsibility to follow any significant leads in the interview, and to be aware of any tendencies on their own part to avoid distressing topics. Experience shows that clinicians tend to be vulnerable to this type of error in particular with regard to sexual functioning, substance use, and racial/ethnic discrimination issues. It may be helpful to keep in mind that while the interview shares many commonalities with social conversation, it is by definition not a run-of-the-mill social interaction. Thus, inhibitions that prevent the interviewer from querying these admittedly uncomfortable topics must be dealt with. Because many clinicians may find themselves having completed much of their formal training without ever overcoming the discomfort experienced when such topics are broached, it may be necessary to practice on colleagues in role-play activity designed to help the clinician become adept at obtaining the necessary information despite initial resistance from within as well as from the respondent. As one of the primary purposes of the clinical interview is accurate diagnosis according to current syndromal criteria from the *Diagnostic and statistical manual of mental disorders* (4th ed., *DSM-IV*), the clinician must have a solid working knowledge of the criteria for major disorders. Many of the diagnostic categories require precise time qualifiers, so any reports of significant symptoms should be followed by the clinician's efforts to establish their time of onset, duration, and severity. The respondent should be encouraged by the clinician to employ descriptive terms, and to indicate in some way the intensity of the symptoms with a numerical scaling or comparative descriptors.

4.04.1.8 Closing the Interview

As the time for the interview draws to a close, the clinician should consolidate the information gained. It is helpful to review one's notes so that any lingering questions may be answered, and to clarify any dates, names, or other details that may have become confused during the course of the interview. An additional responsibility of the clinician conducting the interview is to assist the informant in achieving closure. Many times the clinical interview results in emotional dilation and some level of cognitive disorganization as distressing events are recalled and exposed to another person. The skilled clinician will continue to structure the interview with reminders about the amount of time remaining, summarizing the information provided, and giving appropriate feedback to the informant regarding what to expect next. Avoid rushing the informant out of the room, but be prepared to set limits about the closing of the interview.

When possible, it is beneficial to give the informant a good idea of your diagnostic formulation, and to outline possible intervention strategies. If this is not possible or appropriate at the close of interview, convey to the informant what steps will be taken to complete the evaluation, or provide the informant with an idea of how the information provided will be utilized. If possible the informant should leave the interview with the feeling that they have been heard, understood, and will be benefiting in some way from having participated.

4.04.1.9 The Collateral Interview

Collateral interviewing refers to any direct interviewing done with persons other than the identified patient. Common collateral individuals who are interviewed in the clinical setting include parents, spouses, siblings, and other close relatives. In the case of children and adolescents, school teachers, administrators, and counselors are also often interviewed directly about the behavior and adaptive functioning of the patient. The same skills used in interviewing the identified patient will be employed in these interviews. Empathy, a lack of criticism, and an appropriate use of humor are just as indispensable in talking with a spouse or school principal as they are with the individual presenting for assessment and/or treatment. In many cases, the collateral interview is conducted because the patient is unable to provide the needed information on their own because of disorganizing pathology or other limiting factors, making a collateral information source

even more important. In conducting the collateral interview, one must also determine, to the extent possible, the degree of reliability or weight to place upon the information thus gathered.

The clinician should consider the amount of time the informant has known the patient, and the circumstances under which the patient has been observed by the informant. In the case of the school teacher, beginning with questions regarding the amount of time spent with the patient and the subjects taught provides an opportunity to gather useful information about the school setting. In addition, this allows the clinician to evaluate to some extent the affective responses of the informant toward the patient, for example, excessive anger or frustration on the part of the teacher may point to possible distortions in reporting. It is helpful to probe gently for the teacher's experience level, to avoid being unduly influenced by the observations of one who has relatively little comparative knowledge of normative classroom behavior. Begin by asking how long they have been in this particular school, or teaching this particular grade. If the teacher is special education certified, ask how long they have been certified and in what areas. Usually these queries will be sufficient to obtain a good estimate of the experience base of the teacher, and most will actually respond to the general probes with more than enough information to make appropriate judgments.

In the case of the parent interview, take care to establish current custody arrangements and responsibilities as clearly as possible. Depending upon the jurisdiction in which the clinician works, noncustodial parents may not have the right to seek mental health services for the child. It is the clinician's responsibility to be aware of all the legal constraints on service, as well as the ethical duties peculiar to working with children or others who are unable to consent to treatment legally. Be cautious about taking the word of one parent involved in a visitation or custody dispute who reports that the other parent has no interest in the child, or that the other parent would be completely uninterested in assisting in assessment or treatment for the child. Experience indicates that while this may be true in some cases, this attempt to shut the other parent out of clinical work may result in significant distortion of the presenting facts, and can hamper effective work with the child. Thus, if the parent bringing the child for services indicates that their (ex)spouse will not participate in the interview, go ahead and obtain consent from the present parent to contact the reportedly uninvolved parent. This action would only be contraindicated if the clinician

has convincing evidence that contacting the other parent would present a significant danger to the patient.

4.04.2 DEVELOPMENTAL CONSIDERATIONS IN INTERVIEWING

4.04.2.1 Interviewing Children (Preschool Age through Older Elementary)

Because children are usually brought into the clinic setting by their parents, clinicians typically schedule an interview with the parents to obtain information about current concerns and past history. Parents are in a unique position to provide a chronology of significant events in the child's life, leading up to the present concerns and reasons for referral. Often collateral interviews will be scheduled with others who play a significant role in the child's life, such as grandparents, teachers, day care providers, etc. Indeed, some diagnoses (such as attention deficit hyperactivity disorder [ADHD]) require that symptoms be documented across at least two settings, and it is helpful to have informants from settings such as school to add to the history provided by parents. One should not limit interviewing to only the adults who are significant in the child's life, however. To do so would create a risk of overlooking important information that could be obtained from the child directly about the child's perceived fears, anxieties, mood, and critical events in the child's world. The child's perspective is often overlooked in situations where the child is not articulate about feelings or is immature in language development. It is necessary for the clinician to develop skill in obtaining interview information from children even in these circumstances. An excellent resource for interviewing or observing children, including infants, can be found in Sattler (1998, pp. 96–132).

4.04.2.2 Interviewing Parents

The purpose of the interview with parents is similar to that discussed earlier in the chapter, in that the clinician attempts to clarify the reasons for concern, identify strengths and weaknesses that moderate the presenting problems in the child, and obtain information that could assist with treatment planning. However, there are important ecological variables that are salient for children and should be addressed in the interview. These include placing the child's current and past problems into a social and developmental context, assessing possible risk and resilience factors that may relate to the

child's problems, and assessing the consequences or developmental impact of the child's problems on their future development.

4.04.2.3 Social Context

Schroeder and Gordon (1993) outlined several steps in assessing the problems of young children, including clarifying the referral questions and determining the social context of the problem. Parents often present to clinicians feeling anxiety and/or frustration about their child's problems. This may lead to emotionally-laden, imprecise descriptions of behavior (e.g., "He never minds!" or "She is always disrespectful to her parents!"). The first task in the interview is to help parents define the specific behaviors that cause concern, and to obtain information about the frequency, intensity, and nature of the problem. For instance, a three-year-old child who displays temper tantrums once per week may be of mild concern, but, one who has tantrums three to five times per day would be of much greater concern. The intensity of the child's problems might be gauged by the degree of distress caused to the child or the disruption to typical family activities. For instance, tantrums that occur occasionally at home may cause less distress than if they occur with regularity at church, school, or in other public places. Finally, the nature of the child's problems will be an indicator of severity. Children who engage in cruelty to animals or other people, who are destructive, or who engage in a pattern of fire-setting behavior with the intent to destroy property are of more concern than those who have less serious oppositional and defiant symptoms. As clinicians interview parents about the specific behaviors of concern, important information about the frequency, nature, and severity of the problems can be assessed.

The social context is best assessed by asking simple questions such as, "Who is concerned about the child?," "Why is this person concerned?," and "Why is this person concerned now vs. some other time?" (Schroeder & Gordon, 1993). Although parents or teachers may refer children for assessment or treatment, this does not mean that the child necessarily has a problem that needs treatment. A teacher who refers several active children from a first grade class may be feeling overwhelmed by the sheer number of active children in the class at one time, although a given child's behavior may not be severe enough to warrant a diagnosis of ADHD. Rutter and Schroeder (1981) provided a case of example of a mother who presented with concerns about her daughter occasionally masturbating while watching television. In an

attempt to determine why this mother was concerned and the best approach to intervention, the clinician asked about the mother's perception of what this behavior means. The mother responded by saying that she knew her daughter was at a developmental age when exploring her body was normal, and she knew that nothing bad would happen (e.g., such as growing hair on the palms of her hands) as a result of masturbation. The additional question ("Why is the mother concerned now vs. any other time?") yielded the most salient information about the mother's concerns. The mother revealed that her mother-in-law was coming for a visit the next week, and she was concerned that this relative would have a negative reaction to seeing her granddaughter masturbate. The intervention was simplified by understanding the true reason for the mother's concern. The clinician recommended that the mother provide rules about when and where it was acceptable to masturbate (e.g., when her daughter was alone, in her bedroom, or in the bathroom) and institute a behavioral reward system for remembering not to masturbate while watching television.

Other social contextual information can be obtained about family status (who is living in the home), recent transitions (moves, job changes, births, recent deaths, or illnesses of significant family members), and other family stresses (marital problems, financial stresses, etc.). The presence of persons who are supportive to the child, or who may provide a buffer in the face of other stresses, is important. The literature on resilience is replete with examples of children who have lived with adversity but have developed and functioned normally due to protective factors in their social history (Routh, 1985). The interview with parents also provides an opportunity for assessing possible psychopathology in the parents, such as significant depressive or anxiety symptoms; problems with anger management and self-control, as is often seen in abusive parents; substance abuse problems that may lead to parental instability or neglect; or problems with reality testing, as in schizophrenia. One mother, for example, described her 14-year-old son as being afraid of the dark and reporting seeing ghosts at night. This was viewed by the clinician as an example of a fear that was developmentally inappropriate for a 14-year-old; it also raised questions about possible hallucinations. The context became more clear when the mother revealed that she saw nothing inappropriate about this behavior. The mother reported that she, too, needed to sleep with a light on due to her fear of the dark, and that she also imagined seeing ghosts in her bedroom. This mother reported that she and her

son had many discussions about their mutual fears. The context of the son's fears was changed by the mother's revelation, and the clinician decided to include a more thorough interview regarding the mother's mental status in this case. Even when no concerns about parental psychopathology exist, parental stress levels and affect must be considered when interpreting their reports about child behavior. A parent who is calm and rational in providing a history of their child's behavior may be viewed as more objective than a parent who is extremely upset, tearful, or angry and uses exaggerated descriptors of the child's behavior.

4.04.2.4 Developmental Context

Developmental context provides an essential lens from which to view children's behavior, allowing the clinician to evaluate the child's behavior relative to that of other children of the same chronological and/or mental age. For instance, enuresis may not be unusual in a four-year-old, but would be of concern in a 14-year-old. Likewise, enuresis may not be unusual in a six-year-old youngster with a moderate degree of mental retardation. Some behavioral problems of young children are transient, reflecting their responses to normative developmental challenges (e.g., a five-year-old girl who displays a regression to thumb-sucking and infantile speech patterns following birth of a new sibling). Other problems are more serious and persistent, and suggest risk for later maladjustment. Familiarity with developmental theory and the rich empirical literature in clinical child psychology and developmental psychopathology can provide the clinician with guidance in making these discriminations.

Knowledge of the sequence and transitions in social/emotional development are helpful to the clinician in judging the appropriateness of children's behavior at various ages. For instance, a toddler who has never displayed a strong attachment to a primary caregiver (usually a parent or parents) and who seems to form attachments indiscriminately with others would raise concerns about possible attachment relational problems. A seven-year-old child who cannot delay gratification or consider the feelings of others would be of concern for not having learned appropriate self-control and capacity for emotional empathy that would be expected at that age. Critical developmental tasks for the preschool age child include establishing effective peer relations (e.g., learning to share material resources and adult attention with peers, establishing reciprocal play relationships) and developing flexible self-regulatory skills (e.g., adjusting to the authority

of preschool or daycare teachers and classroom routines). In contrast, children in middle to late elementary years (seven to 12 years of age) encounter developmental tasks related to mastery of knowledge and intellectual skills, leading to feelings of productivity and competence. Children with learning disorders or other developmental problems that interfere with academic progress may be at risk for secondary behavioral or emotional problems related to their primary problems with learning during this developmental period. The clinician must tailor the interview to exploration of the child's strengths and weaknesses in the context of appropriate developmental expectations for particular ages.

The newly emerging field of developmental psychopathology has provided a theoretical and empirical base for better understanding the developmental precursors of psychopathology in children, and the impact of this psychopathology on subsequent functioning (cf, Cicchetti & Cohen, 1995a, 1995b). There is a growing body of research addressing risk factors for the onset and continuity of various childhood disorders. For example, Loeber and colleagues have made important contributions to understanding the developmental pathways to childhood disruptive behavior disorders, in which different constellations of risk factors lead to different outcomes. In their longitudinal study of inner city boys at ages seven, 10 and 13, they found that initiation into antisocial behavior was predicted by some factors (e.g., poor parent-child relations, symptoms of physical aggression) that were present across all three ages, while others (e.g., shyness at age seven, depression at age 10) were age specific (Loeber, Stouthamer-Loeber, Van Kammen, & Farrington, 1991). Further, the environments of children who remained antisocial differed from those whose antisocial behavior dropped out; good supervision was more important in helping older children (age 13 at intake) while attitude toward school was more important for the younger children. Studies such as these illustrate the importance of understanding the contextual variables related to parenting style and parent-child relational issues, as well as specific child behaviors, in determining the significance of presenting problems and their possible trajectory over time.

4.04.2.5 Direct Interview of Children

Perhaps the best and most comprehensive resource guide for interviewing children and adolescents who present with a variety of problems is Sattler's (1998) book on clinical and forensic interviewing of children. Basically,

the goals of the initial interview of the child depends upon the referral questions as well as the age and verbal ability of the child (Sattler, 1998). When interviewing children and their families the information sought often includes the following:

(i) to obtain informed consent to conduct the interview (for older children) or agreement to be at the interview (for younger children);

(ii) to evaluate the children's understanding of why they are at the interview and how they feel about being at the interview;

(iii) to gather information about the children's perception of the situation;

(iv) to identify antecedent and consequent events related to the children's problems;

(v) to estimate the frequency, magnitude, duration, intensity, and pervasiveness of the children's problems;

(vi) to identify the circumstances in which the problems are most or least likely to occur;

(vii) to identify potentially reinforcing events related to the problems;

(viii) to identify factors associated with the parents, school, and environment that may contribute to the problems;

(ix) to gather information about the children's perceptions of their parents, teachers, peers, and other significant individuals in their lives;

(x) to assess the children's strengths, motivations, and resources for change;

(xi) to evaluate the children's ability and willingness to participate in formal testing;

(xii) to estimate what the children's level of functioning was before an injury; and

(xiii) to discuss the assessment procedures and possible follow-up procedures. (Sattler, p. 98).

A part of the interview process with children includes observation of parent and child and obtaining collateral information from the schools or others if the presenting problem relates to learning or behavior problems outside the home. Recognizing the developmental tasks that children must master at varying ages helps the clinician understand the child's behavior. Thus, a comprehensive, detailed developmental history of the child and family milieu is an integral part in establishing an appropriate treatment.

Clinicians must also consider interviewing the child at some stage during the evaluation process. Very young children may be observed using a free-play setting and using observational guides during the play. The clinician can learn a great deal about the child's energy level, physical appearance, spontaneity, organization, behavior, affect, and attitude through their play and through a diagnostic play interview.

School-aged children are able to share thoughts and feelings with the clinician unless they are unusually shy or oppositional (Sattler, 1998). Obviously establishing rapport and maintaining the child's cooperation during the interview is crucial. Kanfer, Eyberg, and Krahn (1992) identified five basic communication techniques that can aid the clinician in attaining rapport and cooperation. First, the clinician can use descriptive statements to describe the clients ongoing behavior, for example, "You're stacking the toys so nice." Second, using reflective statements to mirror the child's statements can be nonthreatening. For example, if the child says she wants to play with blocks the clinician merely reflects "you want to play with the blocks." Third, labeled praise helps the child feel good and feel that the clinician approves of them. Fourth, the clinician must avoid critical statements that suggest disapproval or make the child feel as though they are bad. Finally, open-ended questions avoid yes or no answers and provide opportunities for children to elaborate on their responses (Kanfer et al., 1995).

4.04.2.6 Adolescents

Interpersonal style may play a greater role in good interviewing with this age group than with any other. Adolescents tend to be intensely attuned to any communications that concern their personal appearance, skills, or competence, and the interviewer must avoid at all costs even the hint of condescension. As numerous authors have pointed out, older clinicians tend to identify readily with the parents of adolescents, while younger ones may easily align themselves with the youth. The clinician who remains unaware of their tendencies in this regard runs the risk of making insensitive or intrusive statements that will inhibit rapport rather than increase it. In the first case, the clinician who approaches the adolescent with a parental attitude may unconsciously interact in a way that increases the informant's anxiety, guilt, and hostility. Questions that presuppose information the adolescent has not provided may mirror intrusive interactions with other adults, resulting in defensive efforts and guardedness. Similarly, clinicians who identify easily with the adolescent may also appear "hokey" and insincere when they misuse popular language, or try too hard to relate their own somewhat misty adolescent experiences to those of the youth they are interviewing. These errors result from incautious use of the same techniques that will be necessary for successful adolescent interviewing. That is, to obtain good information and develop adequate rapport, the adolescent must perceive that the

clinician is clearly on their side within the boundaries of the relationship. Judicious use of self-disclosure can help the adolescent believe that the interviewer is not attempting to take away from the interaction without reciprocating. Earnest discussion of the limits of confidentiality and the purposes of the interview will help allay some of the suspicions the informant may have about the clinician's role, and will serve to make a distinction between the clinician-informant relationship and those the adolescent has with parents, teachers, parole officers, and other adults.

The adolescent patient presents a number of challenges to the interviewer that are often less present or significant in interactions with both older and younger people. Because of the unique developmental pressures and challenges of adolescence, special care must be taken in the interview to ensure adequate cooperation as well as to make the interview process a helpful one to the patient. It is essential that the interviewing clinician possess a basic knowledge of the common demands and urges present in the adolescent and their family to effectively assess the patient's functioning. Listed next are those tasks commonly believed to be operating in the adolescent period of life according to various developmental theorists (Erikson, 1963; Rae, 1983).

4.04.2.6.1 Separation-individuation

Separation-individuation refers to the need of the adolescent to identify those qualities in themselves that set them apart from their family. Many of the issues bringing adolescents to treatment involve conflicts that are direct results of this process. The adolescent during this time begins testing family boundaries and experimenting with beliefs and behaviors that differ from those held by their caretakers. This process often produces considerable anxiety for all family members, and the adolescent's interpersonal relations may become quite variable. Often, the adolescent moves between the poles of autonomy from, and dependence upon, the family. An important portion of the adolescent interview is that of identifying the severity of the stressors resulting from this natural process.

4.04.2.6.2 Resolving conflict with authority figures

Related to the individuation task is the frequent occurrence of conflict with authority figures outside of the family as well as within. For younger adolescents this involves primarily their teachers and other school personnel, and

for later adolescents this includes work supervisors as well. Conflicts with authority figures outside the home often have their roots in greater-than-average difficulties in resolving the family relationship struggles. Thus, when interviewing the adolescent, it is helpful to identify both positive and negative relationships with other adults in their life.

Often classroom performance for the adolescent presenting for services is related strongly to the quality of the relationship with the teacher, so discussion of academic performance (usually a relatively nonthreatening issue in the context of the clinical interview) can elicit useful information about this area of functioning as well. Adolescents, as well as younger children, may readily express relational difficulties in response to the question "Is he/she a good teacher?" This often elicits the adolescent's opinion regarding the desirable qualities in an important adult, and allows the interviewer to follow up with questions regarding the adolescent's ability to recognize their own role in any positive or negative interactions.

4.04.2.6.3 Peer group identification

As adolescence is inarguably a time of shifting focus from family relations to peer relations, it is vital to gather information regarding the patient's friendships and any identification with a social subgroup. Some effective ways of eliciting this information include discussion of music topics, such as taste and dress, that will provide clues to the adolescent's social presentation and degree of inclusion or exclusion from social groups.

To effectively interview adolescents regarding social issues, it is necessary for the clinician to maintain a moderate degree of understanding of popular culture. Thus, one would be well served by making an effort to watch television programming, read magazines, and spend time taking in the various electronic media that are aimed at people in this age group. The interviewer should not attempt to present as an authority on the adolescent's culture, but will benefit from being able to recognize specific music groups, current movies, video games and Internet activities, and other elements that are part of the adolescent's milieu. It is often helpful to enlist adolescents' aid in delineating the social groups present in their school, then ask them to identify the group to which they feel they most belong. This question can usually be asked rather directly, and many teens are pleased by the opportunity to display their understanding of the social complexities in their school.

Follow-up inquiry should establish with whom the adolescent spends most time and

how they see themselves as fitting into the groups at school. Many youth social strata include a group delineated primarily by drug/alcohol use as well as different groups for aggressive or delinquent behavior that may be gang-affiliated or gang-emulating. Thus, the social categories to which the adolescent assigns themselves may also point the interviewer toward necessary inquiries into these possible problem areas as well as providing information about the degree of social integration in the adolescent's life.

4.04.2.6.4 Realistic appraisal and evaluation of self-qualities

As the focus of evaluation or treatment is likely to include assessing and modifying self-image, it is necessary to include questions regarding the ways in which the adolescent views themselves. Adolescents generally display both overly optimistic and excessively pessimistic appraisals of personal qualities. One purpose of the interview is to assist in determining when these perceptions are faulty and result in impaired functioning. It is often helpful to present questions about self-image in terms of strengths and liabilities, and to follow up on both. Questions about the adolescent's physical capacities as well as social and emotional abilities are necessary components of the interview. This portion of the interview can be directed toward uncovering problems with perception of body image and behaviors related to physical health. The interviewer should attend carefully to clues that might indicate the need for more focused exploration of possible eating disorders, and to somatic complaints indicative of anxiety or depression.

4.04.2.7 Interviewing Young Adults (18–40 Years)

The psychological distinction between adolescence and young adulthood is frequently blurred, and many of the same traits and problems may be observed in individuals both over and under the chronological age of majority. However, since the age of majority is generally 18 years, a higher proportion of patients over age 18 will be self-referred and hence will present in a more open and cooperative manner than some adolescents. Additionally, young adults are more likely to present with some subjective description of their distress and their situation. Therefore, the client may be more likely to identify a problem area spontaneously. Despite the fact that more patients in this age group may independently seek services, many of the adolescent issues

related to establishing an autonomous role-identity may surface in the interactions with the interviewer, especially with the "youngest" adults. Therefore the interviewer may frequently call upon the skills used in interviewing adolescents.

Erikson (1963) identified the primary developmental conflict for the various stages of adulthood, and these stages suggest important interview topics (see Table 1). The primary conflict of young adulthood is intimacy vs. isolation. Consequently, many of the psychological problem areas frequently encountered will revolve around commitment to interpersonal relationships and establishing trust. Establishment of a working relationship with the patient is also affected by these issues.

A relatively greater amount of the interview might be devoted to exploration of existing relationships or those the patient wishes existed. One type of relationship to consider is that with parents and family of origin. Establishing the degree of desired independence continues to be an issue with some young adults. Issues relevant to these ties might be financial (e.g., parents may be paying college expenses), or they may be more interpersonal in nature (e.g., parents controlling social relationships or defining goals for the patient).

Intimate relationships with individuals of the same or opposite sex may also be a source of psychological discomfort and play a part in the development of anxiety disorders or depression. Inquiry about social functioning should include peer relationships, such as partners in love relationships, friends, and acquaintances.

Individuals in the young adult age group generally will have established some degree of independence, and the relative importance of work and employment will be much greater than at younger ages. The interview should therefore include specific inquiry into current job status, job satisfaction, goals, and relationships with co-workers. The further one progresses into this stage, the greater is the importance of establishment of a stable intimate relationship and mutual trust, and the higher the probability that the issue of procreation will arise. Therefore inquiry should include questions about intentions and concerns associated with having children and child rearing and any differences with one's partner about children.

Finally, the initial episodes of many severe psychiatric disorders are most likely to occur within the young adult period. Initial episodes of depression, and post-partum depression, are likely to occur in those affected before they pass through this period (Kaelber, Moul, & Farmer, 1995). Therefore screening for affective disorders should be included in the interview. A

Table 1 Interview topics for each developmental stage.

Young adult	Independence from family, relationships with peers, stable intimate relationships, trust in relationships, establishment of a family, issues related to having and rearing children, education, and career goals.
Middle adult	Achievement of work and family goals, career or family role changes, responsibility for aging parents, death of grandparents and parents, reducing responsibility for children, changes in physical appearance and characteristics, and anticipating retirement.
Older adult	Accepting status of family and career, developing identity as grandparent or “elder advisor,” coping with reduced physical capability and/or health changes, specific plans for retirement, loss of siblings, spouse, and friends. Increased reliance on children or caretakers.
Late adult	Coping with deteriorating health, decreased mobility, dependence on caretakers, and anticipation of death.

later section of this chapter deals with interviewing depressed and anxious patients. Additionally, first episodes of schizophrenia or bipolar disorder generally take place in adolescence or young adulthood and the interviewer should be sensitive to symptoms of these disorders.

4.04.2.8 Interviewing Adults in Middle Adulthood (40–60 Years)

Interview techniques need not differ with this age group, but the relevant topics from a developmental perspective are somewhat different (see Table 1). This period encompasses much of the creative and productive portion of the life span in western culture. The emphasis is not on starting, but on completing tasks begun in young adulthood. The focus of individuals at this stage of life is much less on goal setting than on goal attainment. The growth and nurturing of an established family, the attainment of successive career goals, and nurturing of one's parents and grandparents occur in this time span. One's children come into adulthood and begin to establish their identities and families. Inquiry into the relationships with the former and succeeding generations should be made. Towards the middle of this period, individuals are able to anticipate the likelihood of reaching family and career goals, and become aware of the fact that certain goals for themselves and their children may not be met. Biological changes associated with mid-life, which are well-defined for women, but also may be present for men, should be queried since they may be associated with depression or anxiety. Possible mid-life existential crises related to loss should also be assessed. The losses may result from death of parents or grandparents, or changes in roles as parent, spouse, or worker.

4.04.2.9 Interviewing Older Adults (60–70 Years)

For many adults in this age range, the predominant life circumstance deals with additional impending changes in the area of life roles. Retirement usually occurs within this time frame, and inquiries might reveal difficulties in psychological adjustment to one's own retirement or the retirement of a significant other. The frequency of death in the patient's social circle gradually increases, and may include a spouse, close friends, or even an adult child.

Due to the possibility of some early decline in cognitive capacity in this age group, the response to inquiry may be defensiveness and denial. The patient with some early impairment may deny the need for the evaluation, object to questions, and become resentful if the interview serves to demonstrate difficulties with memory. Therefore, it becomes more important to interview a collateral person or include a collateral person in the patient interview. In addition to a spouse or family member, a collateral person to be considered with older adults is an adult caretaker, who may or may not be related to the patient. This may give rise to some special issues of confidentiality.

Attention to the collateral person's nonverbal behavior may sometimes suggest that they are uncomfortable reporting the patient's difficulties, especially in the patient's presence. In such circumstances a separate collateral interview is desirable.

4.04.2.10 Interviewing In Late Adulthood (70 Years–End of Life)

Adults in the latest stages of life have their own unique set of circumstances of which the interviewer must be aware. The losses that may have begun earlier may become more frequent.

Physical changes, often represented by medical problems, may interfere with some life activities, and there may be a need to accept reduced independence. At some point anticipation of the end of life is common. The combination of these forces often lead the elderly to have a perspective on life and the situation giving rise to the interview that differs considerably from younger adults, in that they may be unconcerned and see no need for the evaluation. Often the reasons for the interview are more important to someone else than to the patient. As with children and adolescents, it is more likely that someone other than the client identified the need for and arranged for the mental health contact.

It is also common for the oldest adults to answer questions more slowly, either because of difficulty accessing information or because a more tangential and elaborate route is taken to reach a point in conversation. Patience on the part of the examiner in these situations is important, both for maintaining rapport and to show the proper respect due the patient.

It has been estimated that the incidence of cognitive decline in people over age 65 is 10–20% (Brody, 1982). Estimates are as high as 25% of those 80 years and older (Hooper, 1992). Thus, the likelihood of cognitive impairment is even greater in this age group than those discussed previously. For those with cognitive dysfunction, cooperation may be minimal and denial, and even belligerence, may be present. Again, the availability of a collateral person for interview may be very important, as the patient may not cooperate or may be impaired in their ability to provide information.

4.04.3 INTERVIEWING SPECIAL POPULATIONS OR SPECIFIC DISORDERS

4.04.3.1 Interviewing Depressed Patients

Interviewing depressed adults may require some adjustment in the tempo and the goals of the interview. Due to low energy and psychomotor retardation, it may not be possible to gather all the desired information within the time available. Hence, some prioritization of information is necessary, so that issues such as suicidality, need for hospitalization, and need for referral for medication may be addressed. Beck (1967) and later, Katz, Shaw, Vallis, and Kaiser (1995) pointed out that the interpersonal interaction with the depressed patient may be frustrating for the interviewer, not only due to the slowness mentioned above, but also because of the negative affect and negative tone of information provided.

It is also particularly important with depressed patients, who are prone to hopelessness, to provide encouragement and attempt to impart hope to the patient during the interview. This may be done by recognizing areas of strength, either in terms of personal qualities or successful areas of functioning.

Specific inquiry is necessary to diagnose depression appropriately, and a variety of sources are available to guide this inquiry. Diagnostic criteria for depression are clearly delineated in the *DSM-IV* (American Psychiatric Association [APA], 1994). A number of structured interviews have been developed that may serve as guides for inquiry or provide sample questions. Formal training is required for the reliable use of these interviews for diagnostic purposes. The Schedule for Affective Disorders and Schizophrenia (SADS; Endicott & Spitzer, 1978) is a relatively early forerunner of current interviews that slightly preceded the *DSM-III* (APA, 1980), and includes probe questions for depressive symptoms as well as other disorders.

The Structured Clinical Interview for *DSM-III-R* (SCID; Spitzer, Williams, Gibbon, & First, 1992) is a more current instrument with a modular format so that sections for each disorder may be used independently. Table 2 also lists sample questions that might be used to probe for the presence of various depressive symptoms.

4.04.3.2 Interviewing Anxious Patients

The anxious patient may also present some special difficulties during the interview. If the patient is acutely distressed at the time of the interview, as might be true of someone with a generalized anxiety disorder, they may provide a rush of disorganized information so that it may be difficult to obtain a coherent history. Anxiety interferes with attention and concentration, so that repetition may be necessary. Experience has shown that in such a situation, some initial intervention using brief relaxation techniques, is helpful before proceeding with the interview. Anxious patients also frequently seek reassurance that treatment will be effective in reducing their anxiety. It is appropriate to indicate that treatment techniques have been helpful to other anxious patients, and that these techniques will be available to them.

The diagnostic symptoms of various anxiety disorders are identified in *DSM-IV*, and the structured interviews mentioned earlier also provide some guidance for the inquiry for specific anxiety symptoms. In addition to the diagnostic information, it is important to

Table 2 Sample questions for depressive symptoms.

Mood (depressed)	How would you describe your mood? Have you been feeling down or sad much of the time?
Mood (irritable)	How much of the time do you feel down or sad? Have you been more short-tempered than usual for you? Do others say you are more irritable or lose your temper more easily than usual?
Interest and pleasure	Are you as interested as ever in things like your work, hobbies, or sex? Do you continue to enjoy the things you usually like to do, like hobbies, doing things with friends, or your work? Has your interest declined in things which used to be really interesting for you?
Energy/fatigue	Do you have enough energy to do the things you want to do or need to do? Do you have the energy to do the things you find interesting?
Weight loss/gain	Do you tire out more easily than usual? Have you gained or lost weight since . . . (specify a time period)? If the patient does not know, you may inquire about whether clothes fit properly, or what others may have said about weight. Insomnia/hypersomnia
Insomnia/hypersomnia	How well are you sleeping? Do you have difficulty getting to sleep? (initial insomnia). Do you awaken frequently during the night and have trouble getting back to sleep? (middle insomnia)
Psychomotor agitation/retardation	Do you awaken too early in the morning? (terminal insomnia) Have other people commented on your being too active or being very slowed down? Are there times when you just can't sit still, when you have to be active, like pacing the floor or something similar? Are there times when you are very slowed down, and can't move as quickly as usual?
Worthlessness/guilt	How do you feel about yourself? Do you think of yourself as worthwhile? Do you often feel guilty or have thoughts of being guilty for something? Is guilt a problem for you?
Concentration/decisiveness	Is it difficult for you to keep your attention on things you are doing? Do you lose track of things, like conversations or things you are working on? Is there a problem with making decisions? Does it seem that your thoughts are slowed down, so it takes a long time to make a decision?
Thoughts of death/suicide	Do you frequently have thoughts of death? Do you think a lot about friends or loved ones who have died? (Inquire if someone close to the patient has recently died or is near death.) Do you sometimes think it would be better if you were dead? Have you thought about hurting yourself or killing yourself? Have you planned a particular way in which you would kill yourself? What would keep you from killing yourself?

inquire about ways the patient has attempted to cope with the anxiety, and to provide some reinforcement for such efforts.

4.04.4 SUMMARY

The clinical interview provides rich diagnostic information that can aid in the assessment and treatment of patients. Interpersonal style of the

clinician interview, structuring the interview, the setting in which the interview takes place, preparing the patient, and the beginning, middle, and ending phases of the interview are discussed. Developmental considerations and suggestions are offered in interviewing children, adolescents, and adults. Sample questions are primarily for interviewing depressed patients.

4.04.5 REFERENCES

- American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Beck, A. T. (1967). *Depression: Clinical, experimental and therapeutic aspects*. New York: Harper and Row.
- Bricklin, B. (1990). *The custody evaluation handbook: Research-based solutions and applications*. New York: Brunner-Mazel.
- Brody, J. A. (1982). An epidemiologist views senile dementia: Facts and fragments. *American Journal of Epidemiology*, 115, 155–160.
- Cicchetti, D., & Cohen, D. J. (Eds.) (1995a). *Developmental psychopathology. Vol. 1: Theory and methods*. New York: Wiley.
- Cicchetti, D., & Cohen, D. J. (Eds.) (1995b). *Developmental psychopathology. Vol. 2: Risk, disorder, and adaption*. New York: Wiley.
- Corcoran, K., & Vandiver, V. (1996). *Maneuvering the maze of managed care: Skills for mental health professionals*. New York: Simon & Schuster.
- Egan, G. (1994). *The skilled helper: A problem management approach to helping*. Pacific Grove, CA: Brooks/Cole Publishing.
- Endicott, J., & Spitzer, R. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry*, 35, 837–844.
- Erikson, E. H. (1963). *Childhood and society* (2nd ed.). New York: Norton.
- Hooper, C. (1992). Encircling a mechanism in Alzheimer's disease. *Journal of National Institutes of Health Research*, 4, 48–54.
- Kaelber, C. T., Moul, D. E., & Farmer, M. E. (1995). Epidemiology of depression. In E. E. Beckham & W. R. Leber (Eds.), *Handbook of depression* (2nd ed., pp. 3–35). New York: Guilford Press.
- Kanfer, R., Eyberg, S., & Krahn, G. L. (1992). Interviewing strategies in child assessment. In M. Roberts & C. E. Walker (Eds.), *Handbook of clinical child psychology* (pp. 49–62). New York: Wiley.
- Katz, R., Shaw, B., Vallis, M., & Kaiser, A. (1995). The assessment of the severity and symptom patterns in depression. In E. E. Beckham & W. R. Leber (Eds.), *Handbook of depression* (2nd ed., pp. 61–85). New York: Guilford Press.
- Loeber, R., Stouthamer-Loeber, M., Van Kammen, W., & Farrington, D. P. (1991). Initiation, escalation and desistance in juvenile offending and their correlates. *The Journal of Criminal Law and Criminology*, 82, 36–82.
- Luborsky, L. (1996). *The symptom-context method*. Washington, DC: APA Publications.
- Morrison, J. (1995). *The first interview*. New York: Guilford Press.
- Rae, W. A. (1992). Teen-parent problems. In M. C. Roberts & C. E. Walker (Eds.), *Handbook of clinical child psychology* (pp. 555–564). New York: Wiley.
- Routh, M. (1985). Masturbation and other sexual behaviors. In S. Gabel (Ed.), *Behavioral problems in childhood* (pp. 387–392). New York: Grune & Stratton.
- Rutter, D. K., & Schroeder, C. S. (1981). Resilience in the face of adversity: Protective factors and resistance to psychiatric disorder. *British Journal of Psychiatry*, 147, 598–611.
- Sattler, J. (1998). *Clinical and forensic interviewing of children and families* (pp. 96–132). San Diego, CA: J. M. Sattler.
- Schoeder, C. S., & Gordon, B. N. (1993). Assessment of behavior problems in young children. In J. L. Culbertson & D. J. Willis (Eds.), *Testing young children* (pp. 101–127). Austin, TX: ProEd.
- Spitzer, R., Williams, J. B. W., Gibbon, M., & First, M. (1992). The Structured Clinical Interview for *DSM-III-R* (SCID): I. History, rationale and description. *Archives of General Psychiatry*, 49, 624–636.

4.05

Structured Diagnostic Interview Schedules

JACK J. BLANCHARD and SETH B. BROWN
University of New Mexico, Albuquerque, NM, USA

4.05.1 INTRODUCTION	98
4.05.1.1 <i>Evaluating Structured Interviews</i>	99
4.05.1.1.1 <i>Reliability</i>	99
4.05.1.1.2 <i>Validity</i>	101
4.05.1.2 <i>Overview</i>	101
4.05.2 ADULT DISORDERS	101
4.05.2.1 <i>Schedule for Affective Disorders and Schizophrenia</i>	101
4.05.2.2 <i>Reliability</i>	102
4.05.2.2.1 <i>Summary</i>	104
4.05.2.3 <i>Present State Examination</i>	104
4.05.2.3.1 <i>Reliability</i>	105
4.05.2.3.2 <i>Supplements to the PSE</i>	106
4.05.2.3.3 <i>Summary</i>	107
4.05.2.4 <i>Structured Clinical Interview for DSM-IV/Axis I Disorders</i>	107
4.05.2.4.1 <i>Reliability</i>	108
4.05.2.4.2 <i>Summary</i>	109
4.05.2.5 <i>Comprehensive Assessment of Symptoms and History</i>	109
4.05.2.5.1 <i>Reliability</i>	110
4.05.2.5.2 <i>Summary</i>	110
4.05.2.6 <i>Diagnostic Interview for Genetic Studies</i>	111
4.05.2.6.1 <i>Reliability</i>	111
4.05.2.6.2 <i>Summary</i>	112
4.05.2.7 <i>Diagnostic Interview Schedule</i>	112
4.05.2.7.1 <i>Reliability</i>	112
4.05.2.7.2 <i>Summary</i>	113
4.05.2.8 <i>Composite International Diagnostic Interview</i>	113
4.05.2.8.1 <i>Reliability</i>	114
4.05.2.8.2 <i>Summary</i>	114
4.05.3 PERSONALITY DISORDERS	114
4.05.3.1 <i>Structured Interview for DSM-IV Personality Disorders</i>	114
4.05.3.1.1 <i>Reliability</i>	115
4.05.3.1.2 <i>Summary</i>	115
4.05.3.2 <i>International Personality Disorder Examination</i>	116
4.05.3.2.1 <i>Reliability</i>	117
4.05.3.2.2 <i>Summary</i>	117
4.05.3.3 <i>Structured Clinical Interview for DSM-IV Personality Disorders</i>	118
4.05.3.3.1 <i>Reliability</i>	119
4.05.3.3.2 <i>Summary</i>	119
4.05.3.4 <i>Personality Disorder Interview-IV</i>	120
4.05.3.4.1 <i>Reliability</i>	120
4.05.3.4.2 <i>Summary</i>	120
4.05.4 CHILD AND ADOLESCENT DISORDERS	121

4.05.4.1 Schedule for Affective Disorders and Schizophrenia for School Age Children	121
4.05.4.1.1 Reliability	121
4.05.4.1.2 Summary	121
4.05.4.2 Child Assessment Schedule	122
4.05.4.2.1 Reliability	122
4.05.4.2.2 Summary	123
4.05.4.3 Child and Adolescent Psychiatric Assessment	123
4.05.4.3.1 Reliability	124
4.05.4.3.2 Summary	124
4.05.4.4 Diagnostic Interview Schedule for Children	124
4.05.4.4.1 Reliability	124
4.05.4.4.2 Summary	125
4.05.4.5 Diagnostic Interview for Children and Adolescents	125
4.05.4.5.1 Reliability	126
4.05.4.5.2 Summary	126
4.05.5 SUMMARY	126
4.05.6 REFERENCES	126

4.05.1 INTRODUCTION

As early as the 1930s, and certainly by the early 1960s, it was apparent that clinical diagnostic interviews were fallible assessments. Evidence suggested that clinicians frequently arrived at different diagnoses, often agreeing at no more than chance levels (e.g., Beck, Ward, Mendelson, Meck, & Erbaugh, 1962; Matarazzo, 1983; Spitzer & Fleiss, 1974). These findings gave rise to the study of the causes of diagnostic unreliability as well as the development of methods to improve psychiatric diagnosis.

In the first study systematically to examine reasons for diagnostic disagreement, Ward, Beck, Mendelson, and Erbaugh (1962) summarized three major sources of disagreement. These were inconsistency on the part of the patient (5% of the disagreements), inconsistency on the part of the diagnostician (32.5%), and inadequacies of the nosology (62.5%). Thus, nearly one-third of the diagnostic disagreements arose from the diagnosticians. Factors associated with differences between diagnosticians included interviewing techniques that led to differences in information obtained, weighing symptoms differently, and differences in how the symptomatology was interpreted (Ward et al., 1962). It is interesting that these problems arose despite methods of study which included the use of experienced psychiatrists, preliminary meetings to review and clarify diagnostic categories, the elaboration of diagnostic descriptions to minimize differences, and the compilation of a list of instructions to guide diagnosis (Beck et al., 1962).

The study of Ward et al. (1962) identified two major sources of disagreement that have been termed "criterion variance" and "information variance" (Endicott & Spitzer, 1978). Criterion variance refers to the errors in diagnostic

assignment attributable to how clinicians summarize patient information into existing definitions of psychiatric diagnoses. Inadequacies of early diagnostic systems (e.g., the *Diagnostic and statistical manual of mental disorders* [DSM-I and DSM-II]) generally arose from the lack of explicit diagnostic criteria. The development of newer diagnostic schemes such as the Research Diagnostic Criteria (RDC; Spitzer, Endicott, & Robins, 1978), and subsequently the *DSM-III* (American Psychiatric Association, 1980), provided inclusion and exclusion criteria and specific criteria relating to symptoms, signs, duration and course, and severity of impairment.

In addressing errors that arise from inadequate nosology, clinicians and researchers are still faced with information variance, that is, errors originating from differences in what information is obtained and how that information is used by the interviewers. As reviewed above, Ward et al. (1962) found that nearly a third of diagnostic disagreements were related to the interviewers. Structured interviews were developed in order to address this source of error variance, and the history of structured interviews goes back to the 1950s (Spitzer, 1983). All structured interviews seek to minimize information variance by ensuring that clinicians systematically cover all relevant areas of psychopathology. Although specific methods vary across instruments, common techniques that characterize structured interviews are the specification of questions to be asked to assess domains of psychopathology and the provision of anchors and definitions in order to determine the ratings of symptoms (e.g., do the descriptions obtained within the interview achieve diagnostic threshold or not?). Despite some shared qualities it is also clear that available structured interviews differ markedly on a number of dimensions. The reliability of

diagnoses based on refined diagnostic criteria paired with structured interviews was found to be greatly improved (e.g., Endicott & Spitzer, 1978; Spitzer, Endicott, & Robins, 1978).

4.05.1.1 Evaluating Structured Interviews

The selection of a structured interview will be driven by a number of other concerns. Some of the potential considerations are summarized in Table 1 and are derived from the reviews of Page (1991) and Zimmerman (1994). Questions that should be asked in selecting an instrument will include the diagnoses covered by the interview, the nosological criteria adhered to in generating these diagnoses (e.g., *DSM*, the *International classification of diseases [ICD]*, or other criteria such as the RDC), and the population studied (e.g., adult or child). Additionally, the context in which the interview is conducted may also be relevant. That is, who will be administering the questionnaire and under what circumstances? Some interviews were developed to be used by lay interviews as in community epidemiological studies while other instruments require extensive clinical experience and are to be administered only by mental health professionals. Other concerns will relate to the guidelines and support available for an instrument. Some measures have extensive book-length user's manuals with available training videotapes and workshops; however, other measures have only sparse unpublished manuals with no training materials available. Finally, major concerns arise regarding reliability.

4.05.1.1.1 Reliability

The reliability of a diagnostic interview refers to the replicability of diagnostic distinctions obtained with the interview. Methods for evaluating agreement between two or more raters can take a variety of forms (Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981). Importantly, these differing methods may have an impact on indices of reliability. The most stringent evaluation involves raters conducting separate interviews on different occasions with diagnoses based only on interview data (i.e., no access to medical records or collateral informants such as medical staff or family members). Reliability assessment based on this method is referred to as "test-retest reliability," given the two occasions of interviewing. This methodology ensures a rigorous evaluation of the ability of the interview to limit information variance and to yield adequate information for the determination of diagnoses. However, in addition to interviewing style and

methods of diagnosticians, two other factors contribute to rater disagreements in test-retest designs. First, the information provided by the patient may be different during the two interviews. Even with structured interviews this can continue to be a relevant source of variance. In a review of test-retest data using the Schedule for Affective Disorders and Schizophrenia, Lifetime Anxiety, Fyer et al. (1989) found that more than 60% of diagnostic disagreements were due to variation in the information provided by subjects. Second, there may be a true change in the clinical status of the individual who is interviewed. As the test-retest period increases, the potential contribution of changing clinical status will increase.

Other methods sometimes utilize a single interview that is either observed or videotaped and rated by independent (nonparticipating) raters, yielding inter-rater agreement. This method may yield inflated estimates of reliability as information variance is absent; that is, both raters are basing diagnostic decisions on the same subject responses. Also, interviewer behavior may disclose diagnostic decisions to the second rater. For example, during a structured interview an interviewer may determine that a module of the interview is not required based on subject responses and the interviewer's interpretation of rule-out criteria. The observing rater is aware of this diagnostic decision as the interviewer skips questions and moves to another module. Given the importance of methods used in assessing reliability, throughout this chapter we will attempt to indicate clearly the techniques by which reliability was assessed for each instrument.

In addition to considering study designs used in evaluating diagnostic reliability, it is also important to examine the statistics used to compute reliability. One method that has been used as an index of reliability is percent agreement. As outlined by Shrout, Spitzer, and Fleiss (1987), percent agreement is inflated by chance agreement and the base rates with which disorders are diagnosed in a sample. In their example, if two clinicians were randomly to assign *DSM* diagnoses to 6% of a community sample of 100 persons, chance agreement would produce an overall rate of agreement of about 88.8% (Shrout et al., 1987).

The index of agreement, kappa (*K*), was developed to address this statistical limitation (Cohen, 1960). Kappa reflects the proportion of agreement, corrected for chance; it varies from negative values reflecting agreement below chance, zero for chance agreement, and positive values reflecting agreement above chance to 1.0 for perfect agreement (Spitzer, Cohen Fleiss, & Endicott, 1967). The statistic weighted *K* was

Table 1 Relevant questions for selecting a diagnostic interview.*Content*

Does the interview cover the relevant diagnostic system (e.g., RDC, *DSM-IV*, *ICD-10*)?

As an alternative to adhering to a single diagnostic system, does the interview provide polydiagnostic assessment (i.e., diagnoses for multiple diagnostic systems can be generated)?

Does the interview cover the relevant disorders?

Can irrelevant disorders be omitted?

Does the interview provide a sufficiently detailed assessment? That is, aside from diagnostic nosology is adequate information in other domains assessed (e.g., course of illness, family environment, social functioning)?

How are signs and symptoms rated (e.g., dichotomous ratings of presence vs. absence or continuous ratings of severity)?

Population

Is the interview applicable to the target population?

Relevant population considerations include adult vs. child, patient or nonpatient, general population, or family members of psychiatrically ill probands and whether the instrument will be used cross-culturally (is it available in languages for the cultures to be studied?).

Aside from general population considerations are there other exclusionary conditions to be aware of (e.g., age, education, cognitive functioning, or exclusionary clinical conditions)?

Time period

Does the interview cover the relevant time period (e.g., lifetime occurrence)?

Can the interview be used in longitudinal assessments to measure change?

Logistics of interview

How long does the interview take?

Does interview require or suggest use of informant (e.g., with child interviews)?

Administration/interviewer requirements

Who can administer the interview (e.g., lay interviewers, mental health professionals)?

How much training or experience is required to administer interview?

Does interview provide screening questionnaire to assist in expediting the interview (e.g., in personality disorder assessments)?

Are computer programs required and compatible with available equipment?

Guidelines and support

What guidelines for administration and scoring are available (e.g., user's manual)?

What training materials are available (videotapes, workshops)?

Is consultation available for training or clarification of questions regarding administration or scoring?

Reliability and validity

Is the interview sufficiently reliable?

Are reliability data available for the diagnoses and populations to be studied?

Are validity data available for the interview (e.g., concordance with other structured interviews, expert-obtained longitudinal data, and other noninterview measures)?

developed for distinguishing degrees of disagreement, providing partial credit when disagreement is not complete (Spitzer et al., 1967).

Standards for interpreting kappa suggest that values greater than 0.75 indicate good reliability, values between 0.50 and 0.75 indicate fair reliability, and values below 0.50 reflect poor reliability (Spitzer, Fleiss, & Endicott, 1978). Although the present review will adhere to the recommendations of Spitzer, Fleiss & Endicott (1978) that kappas below 0.50 indicate poor agreement or unacceptable reliability, it should

be noted that other authors have proposed what appear to be more lenient criteria for evaluating kappa. For example, Landis and Koch (1977) suggest that kappas as low as 0.21–0.40 be considered as indicating “fair” agreement.

It is important to understand that reliability is not a quality inherent within an instrument that is transportable to different investigators or populations. All the measures described herein require interviewer training, and some instruments require an extensive amount of prior clinical experience and professional training.

Ultimately, the reliability of any structured interview will be dependent on the user of that interview. Although the present review will invite comparisons of reliability across studies it should be noted that reliability statistics such as kappa are influenced by a number of factors that constrain such comparisons. Differences in population heterogeneity, population base rates, and study methods will all influence reliability and should be considered in evaluating the literature.

4.05.1.1.2 Validity

In addition to reliability, the validity of a diagnostic assessment can also be evaluated. In the absence of an infallible or ultimate criterion of validity for psychiatric diagnosis, Spitzer (1983) proposed the LEAD standard: longitudinal, expert, and all data. Longitudinal refers to the use of symptoms or information that may emerge following an evaluation in order to determine a diagnosis. Additionally, expert clinicians making independent diagnoses based on all sources of information make a consensus diagnosis that will serve as a criterion measure. These expert clinicians should use all sources of data that have been collected over the longitudinal assessment including direct evaluation of the subject, interviewing of informants, and information from other professionals such as ward nurses and other personnel having contact with the subject. Typically, validity data such as that suggested by Spitzer are rarely available for structured interviews.

4.05.1.2 Overview

Within this chapter we provide an overview of the major structured interviews available for use with adult and child populations. The interviews included in this chapter are listed in Table 2. Due to space limitations we have focused on the review of broad diagnostic instruments and have not reviewed more narrow or specialized instruments that may address a single diagnosis or category of diagnoses. Each instrument will be reviewed with regard to its history and development and description of the instrument including its format and the diagnoses covered. Reliability data available will be presented and reviewed. Finally, a summary will be provided that intends to highlight the advantages and disadvantages inherent in each instrument. Interviews reviewed will address adult disorders, including personality disorders, and interviews for children and adolescents.

4.05.2 ADULT DISORDERS

4.05.2.1 Schedule for Affective Disorders and Schizophrenia

In order to address sources of diagnostic error arising from criterion variance, Spitzer, Endicott, and Robins (1978) developed the RDC. The RDC contains specific inclusion and exclusion criteria for 25 major diagnostic categories as well as subtypes within some categories. Disorders covered include schizophrenia spectrum disorders, mood disorders (depression and bipolar disorders), anxiety disorders (panic, obsessive-compulsive disorder, phobic disorder, and generalized anxiety disorder), alcohol and drug use disorders, some personality disorders (cyclothymic, labile, anti-social), and two broad categories of unspecified functional psychosis, and other psychiatric disorder. The major source of data for determining RDC diagnoses is the use of the Schedule for Affective Disorders and Schizophrenia (SADS; Endicott & Spitzer, 1978).

As originally developed there were three versions of the SADS: the regular version (SADS), the lifetime version (SADS-L), and a change version (SADS-C). A lifetime anxiety version of the SADS (SADS-LA; Fyer et al, 1989; Mannuzza et al., 1989) has been developed to assess RDC, *DSM-III*, and *DSM-III-R* criteria for almost all anxiety disorder diagnoses, in addition to all the diagnoses covered in the original SADS. The SADS has two parts: Part I provides a detailed description of current condition and functioning during the one week preceding the interview, Part 2 assesses past psychiatric disturbance. The SADS-L is similar to Part 2 of the SADS but the SADS-L focuses on both past and current disturbance. It is appropriate for use in populations where there is likely no current episode or when detailed information regarding the current condition is not required. Endicott and Spitzer (1978) estimate that the SADS can be completed in one and one-half to two hours depending on the disturbance of the individual being interviewed.

The SADS provides probe questions for each symptom rating. However, in addition to using the interview guide the rater is instructed to use all sources of information and to use as many supplemental questions as is required to make ratings. Part 1 of the SADS rates severity of symptoms when they were at their most extreme. Many items are rated for severity during the week prior to the interview and for severity during their most extreme during the current episode. Ratings are made on a seven-point scale from zero (no information) to six (e.g., extreme). The SADS provides defined

Table 2 Structured interviews included in review.*Adult*

Schedule for Affective Disorders and Schizophrenia
 Present State Examination
 Structured Clinical Interview for *DSM-IV*/Axis I Disorders
 Comprehensive Assessment of Symptoms and History
 Diagnostic Interview for Genetic Studies
 Diagnostic Interview Schedule
 Composite International Diagnostic Interview

Personality disorders

Structured Interview for *DSM-IV* Personality
 International Personality Disorder Examination
 Structured Clinical Interview for *DSM-IV* Personality Disorders
 Personality Disorder Interview-IV

Child and adolescent

Schedule for Affective Disorders and Schizophrenia for School Age Children
 Child Assessment Schedule
 Child and Adolescent Psychiatric Assessment
 Diagnostic Interview Schedule for Children
 Diagnostic Interview for Children and Adolescents

levels of severity for each item. For example, in the screening items for manic criteria the item "less need for sleep than usual to feel rested" ratings are as follow: 1 = no change or more sleep needed; 2 = up to 1 hour less than usual; 3 = up to 2 hours less than usual; 4 = up to 3 hours less than usual; 5 = up to 4 hours less than usual; 6 = 4 or more hours less than usual.

In addition to the item ratings and the assignment of RDC diagnoses, the SADS can be used to provide eight summary scales: Depressive Mood and Ideation, Endogenous Features, Depressive-Associated Features, Suicidal Ideation and Behavior, Anxiety, Manic Syndrome, Delusions-Hallucinations, Formal Thought Disorder. These scales were determined by factor-analytic work using similar content scales, and an evaluation of clinical distinctions that are made in research of affective and schizophrenic disorders. The intent of the scales is to provide a meaningful summary of SADS information.

The SADS is intended for use by individuals with experience in clinical interviewing and diagnostic evaluation. Since clinical judgments are required in determining the need for supplemental questioning and in determining ratings, Endicott and Spitzer (1978) suggest that administration of the interview be limited to psychiatrists, clinical psychologists, or psychiatric social workers. However, these authors do note that interviewers with different backgrounds may be used but will require additional training. In one study using highly trained raters (Andreasen et al., 1981), reliability in rating

videotaped SADS interviews was not affected by level of education (from medical degrees and doctorates in psychology to masters and other degrees) or years of prior clinical experience (from less than four years to more than 10 years).

Training videotapes and training seminars are available from the developers of the SADS at the New York State Psychiatric Institute, Columbia University. One example of training in the SADS is provided by Mannuzza et al. (1989) in the use of the SADS-LA. Training was conducted in three phases consisting of 50–60 hours over three to four months. In Phase 1 raters spent 20 hours attending lectures covering diagnosis, systems of classification, interviewing technique and the SADS-LA rater manual, reviewed RDC and *DSM-III* vignettes, and rated videotapes of interviews. In Phase 2 each rater administered the SADS-LA with one patient with an expert trainer and other raters observed. Interviews were subsequently reviewed and discussed. In the final phase, raters independently interviewed three or four patients who had already received the SADS-LA from an expert rater. This final phase allowed for test-retest reliability to be examined and provided an opportunity to discuss discrepancies in ratings during a consensus meeting.

4.05.2.2 Reliability

Initial reliability data for the SADS were reported by Spitzer, Endicott, and Robins

(1978) and Endicott and Spitzer (1978) for both joint interviews ($N = 150$) and independent test-retest interviews, separated by no more than one week ($N = 60$). For joint interviews, present and lifetime RDC diagnoses obtained kappas greater than 0.80 (median $K = 0.91$), with the exception of minor depressive disorder ($K = 0.68$). For test-retest interviews, reliability was somewhat attenuated but remained high with kappas greater than 0.55 for all disorders (median $K = 0.73$) with the exception of bipolar I (0.40). Endicott and Spitzer (1978) also reported reliability of the SADS items and eight summary scales using these same samples. For the 120 items of the current section of the SADS, reliability was high for both joint (90% of items interclass correlation coefficients [ICCs] equal to or greater than 0.60) and test-retest interviews (82% of items ICCs greater than or equal to 0.60). Summary scales also yielded high reliability for joint (ICC range = 0.82–0.99, median = 0.96) and test-retest interviews (ICC range = 0.49–0.91, median = 0.83). Spitzer et al. (1978) also examined the reliability of the SADS-L with first-degree relatives of patient probands ($N = 49$). All kappas were 0.62 or higher with the exception of other psychiatric disorder (0.46), median kappa = 0.86.

Two subsequent studies examined test-retest reliability (separate interviews conducted on the same day) of the SADS (Andreasen et al., 1981; Keller et al., 1981). In a study of 50 patients using the SADS-L, Andreasen et al. (1981) found ICCs equal to or greater than 0.62 for the major RDC diagnoses of bipolar I and II, major depressive disorder, alcoholism, and never mentally ill. The RDC subtypes of major depression also achieved ICCs equal to or greater than 0.60 with the exception of the subtype of incapacitating. Keller et al. (1981), using the SADS-L in a sample of 25 patients, obtained kappas equal to or greater than 0.60 for the RDC diagnoses of schizophrenia, schizoaffective-depressed, manic, major depressive disorder, and alcoholic. The major diagnoses of schizoaffective-manic and hypomanic had low reliability with kappas of 0.47 and 0.26, respectively. Keller et al. (1981) also found high reliability for background information items on social and educational background, and history of hospitalization (kappas greater than 0.73). Finally, individual items from manic, major depressive disorder, psychosis, alcohol and drug abuse, suicidal behavior, and social functional all achieved kappas above 0.56.

McDonald-Scott and Endicott (1984) evaluated the influence of rater familiarity on diagnostic agreement using the SADS. In this

study modified SADS-C ratings were compared for two raters: one with extensive familiarity about subject's psychiatric history and course of illness and prior history and one who was blind to this history and had no prior contact with the subject. Quasi-joint interviews were conducted with the two raters. The nonblind rater was allowed to ask additional questions following the joint SADS interview, in the absence of the blind rater. Of four SADS-C summary scale scores all achieved ICCs of 0.79 or greater. At the item level, 92% of the 52 items had ICCs of 0.60 or greater. Rater differences in scoring suggested that the blind rater may have been somewhat more sensitive to items relating to dysphoria while the nonblind rater was more likely to identify some symptoms that may have been minimized or missed in the blind rater's interview (e.g., mania). However, these discrepancies were subtle and suggest that the SADS can achieve accurate assessment of current cross-sectional functioning whether or not raters have familiarity with the patient.

The inter-rater reliability of the SADS-derived *DSM-III* diagnoses in adolescents has been examined by Strober, Green, and Carlson (1981). Joint interviews were conducted with 95 inpatient adolescents and a family member. Raters independently reviewed all available collateral information prior to the SADS interview (e.g., medical and psychiatric records, school records, current nurses' observations). All diagnoses achieved kappas of 0.63 or greater with the exception of anxiety disorders of childhood (0.47) and undiagnosed illness (0.47). Although encouraging, these data should be viewed in the context of the use of joint interviews and the extensive use of collateral information to supplement the SADS.

Mannuzza et al. (1989) examined the reliability of the SADS-LA in a sample of 104 patients with anxiety disorders. Independent interviews were conducted with test-retest periods ranging from the same day to 60 days. Collapsing across RDC, *DSM-III*, and *DSM-III-R* anxiety disorder diagnoses, agreement for lifetime disorders achieved kappas of 0.60 or greater, with the exception of simple phobia. Examining lifetime anxiety diagnoses separately for each diagnostic system again suggested adequate reliability for most disorders (K range = 0.55–0.91), with the exception of RDC and *DSM-III-R* diagnoses of simple phobia and generalized anxiety disorder (K s less than 0.49). Using this same sample, Fyer et al. (1989) assessed item reliability and factors contributing to disagreements. In general, symptoms were reliably rated with the exception of stimulus-bound panic (typical of simple

phobia), near panic attacks, persistent generalized anxiety, six social and nine nonsocial irrational fears. Review of narratives and consensus meeting forms by Fyer et al. (1989) suggested that the largest source of disagreement was variation in information provided by subject (more than 60% of disagreements). Differences in rater interpretation of criteria resulted in 10–20% of the disagreements and rater error accounted for 10% of the disagreements.

The prior studies have examined test–retest reliability of lifetime diagnoses over brief periods. Two studies have examined the long-term test–retest reliability of SADS-L diagnoses. Bromet, Dunn, Connell, Dew, and Schulberg (1986) examined the 18-month test–retest reliability of the SADS-L in diagnosing lifetime major depression in a community sample of 391 women. Whenever possible, interviewers conducted assessments with the same subject at both interviews. Overall, reliability of lifetime diagnoses of RDC episodes of major depression was quite low. Of those women reporting an episode of major depression at either interview for the period preceding the first assessment, only 38% consistently reported these episodes at both interviews (62% reported a lifetime episode on one occasion but not another). For those women meeting lifetime criteria for a depressive episode at the first interview, fully 52% failed to meet criteria at the time of the second interview.

In a large-scale study of 2226 first degree relatives of probands participating in the National Institute of Mental Health (NIMH) Collaborative Program on the Psychobiology of Depression study, Rice, Rochberg, Endicott, Lavori, and Miller (1992) examined the stability of SADS-L-derived RDC diagnoses over a six-year period. The rater at the second interview was blind to the initial SADS-L. A large degree of variability in reliability was obtained for RDC diagnoses, with kappas ranging from 0.16 to 0.70. Diagnoses with kappas greater than 0.50 included major depression, mania, schizoaffective-mania, alcoholism drug use disorder, and schizophrenia. Diagnoses with low reliability as reflected by kappas below 0.50 were hypomania, schizoaffective-depressed, cyclothymia, panic disorder, generalized anxiety disorder, phobic disorder, antisocial personality, and obsessive-compulsive disorder. Rice et al. (1992) suggested that diagnostic reliability increases with symptom severity. In the studies of Bromet et al. (1986) and Rice et al. (1992), results indicated that there may be substantial error in the temporal stability of some SADS-L-derived lifetime diagnoses. This error may be particularly problematic in nonclinical community samples as studied in these investigations.

4.05.2.2.1 Summary

The development of the SADS represented significant progress in clinical assessment. The SADS has been used extensively in a number of research studies and a wealth of reliability data are available. The SADS provides a broad assessment of symptoms as well as severity ratings for many symptoms. However, the range of disorders covered is somewhat narrow (with an emphasis on schizophrenia, mood disorders, and anxiety disorders in the SADS-LA). Additionally, diagnostic criteria are based on the RDC, with the exception of anxiety disorders covered in the SADS-LA which provide *DSM-III-R* diagnoses.

4.05.2.3 Present State Examination

The Present State Examination (PSE) grew out of research projects in the UK requiring the standardization of clinical assessment. The PSE was not developed as a diagnostic instrument, as with the SCID and SADS. Rather, the PSE was intended to be descriptive and facilitate investigation of diagnostic rules and practices. At the time of the first publication of this instrument (Wing, Birley, Cooper, Graham, & Isaacs, 1967), the PSE was in its fifth edition. Currently, the ninth edition is widely used (Wing, Cooper, & Sartorius, 1974) and the tenth edition of the PSE is available (Wing et al., 1990). The PSE has been translated into over 40 languages and has been used in two large-scale international studies: the US–UK Diagnostic Project (Cooper et al., 1972) and the International Pilot Study of Schizophrenia (IPSS; World Health Organization, 1973).

The standardization of the PSE is achieved through the provision of a glossary of definitions for the phenomena covered by the interview. Additionally, specific series of questions with optional probes and cut-off points are also provided. Detailed instructions for rating the presence and severity of symptoms is also available. Despite this standardization, the developers have emphasized that it remains a clinical interview. The examiner determines the rating provided, evaluates the need for additional probe questions, and uses a process of cross-examination to address inadequate or inconsistent responses. As the name implies, the PSE was developed to ascertain present symptomatology and focuses on functioning in the month prior to the interview.

The eighth edition of the PSE was comprised of 500 items which were then reduced to 140 symptom scores. The ninth edition of the PSE reduced the number of items by having the 140 symptoms rated directly (the presence or

absence of a symptom can be determined without asking as many questions, although additional probe questions are maintained in the ninth edition). Items receive one of three ratings. A zero indicates that a symptom is not present. If present, a symptom is rated as either one (moderate) or two (severe). Items are grouped into symptom scores based on item content and infrequency (Wing et al., 1974). The eighth edition takes approximately one hour to complete while the ninth edition takes approximately 45 minutes (Wing et al., 1974).

Symptoms can be further reduced to 38 syndrome scores by grouping together symptoms of similar content. For example, in the ninth edition the symptoms of worrying, tiredness, nervous tension, neglect through brooding, and delayed sleep are combined into the syndrome score of "Worrying, etc." These syndrome scores were intended to aid in the process of diagnosis by reducing the information to be considered, provide descriptive profiles, and provide a brief method of summarizing clinical information from other, non-interview, sources such as medical records by using a syndrome checklist.

Following the rating of items, a computer program (CATEGO) can be used to summarize PSE ratings further. For the ninth edition, the CATEGO program provides syndrome scores along with summary data for each syndrome (e.g., scores on constituent items). In the next stage, the program further summarizes the syndrome scores into six descriptive categories. The certainty of each descriptive category is also indicated (three levels of certainty are provided). Finally, a single CATEGO class (of 50) is assigned. Importantly, Wing (1983) has emphasized that the PSE and CATEGO program were not developed as diagnostic instruments *per se*. The CATEGO category or class assignments should not be considered diagnoses in the sense of DSM or ICD nosology. Rather, these summaries are provided for descriptive purposes. However, data from the US-UK Diagnostic Project and the IPSS have indicated reasonable convergence between CATEGO classes and clinical project diagnoses, especially when clinical history information is added to the PSE (reviewed in Wing et al., 1974).

Although short training courses lasting one week are available at various centers including the Institute of Psychiatry in London, Wing (1983) suggests that more extensive training is necessary. Wing (1983) recommends that at least 20 interviews be conducted under supervision in order to determine competency in administration of the PSE. Luria and Berry (1980) describe the stages of training used by

these authors to achieve reliable PSE administration. In this study a general introduction and experience with unstructured symptom assessment was followed by reading and discussion of the PSE manual, the rating and discussion of 13 videotaped PSE interviews, and finally, participation in and observation of 12 live student-conducted PSE interviews followed by discussion.

4.05.2.3.1 Reliability

Early evaluations of the reliability of the PSE indicated promising agreement between raters. In the first reliability study conducted on early versions of the PSE (up to PSE-5), rater agreement was evaluated with both independent interviews and observation of interviews, or listening to audiotapes (Wing et al., 1967). Assignment to main categories suggested reasonable agreement, using percent agreement, of 83.7%. Examining five nonpsychotic symptoms, agreement also seemed satisfactory (range across studies $r = 0.53-0.97$). Reliability for nine psychotic symptoms, calculated for single interviews (tape recorded or observed) was also adequate (range of $r = 0.62-0.97$). Kendell, Everitt, Cooper, Sartorius, and David (1968) found a mean kappa for all items to be 0.77.

Luria conducted two reliability studies using the PSE-8 (Luria & McHugh, 1974; Luria & Berry, 1979). Luria and McHugh (1974) examined agreement using six videotaped PSE interviews. The authors examined agreement for 19 profiles of their own design. Patients were ranked on each category based on ratings of examiners. Reliability for these categories was generally adequate with Kendall's W coefficients greater than 0.73 except for behavioral items such as psychomotor retardation (0.66); excitement, agitation (0.47), catatonic, bizarre behavior (0.44); blunted, inappropriate, incongruous affect (0.60). In a subsequent study, Luria and Berry (1979) examined agreement on 20 symptoms deemed of diagnostic importance, 19 psychopathology profiles, and eight syndromes. Thirteen interviews were rated for reliability on the basis of videotapes; 12 were rated based on joint observational interviews. Reliability for videotape and live symptom ratings were adequate with median ICCs of 0.84 and 0.86, respectively (however, agitation or retardation and bizarre behaviors were judged to have poor reliability). Of the 19 profile ratings, the 13 had adequate reliability for videotaped (0.97) and live interviews (0.95). The six behavioral profiles were somewhat lower at 0.72 and 0.66, respectively. Syndrome agreement was high with generalized kappas above 0.91.

Three studies have examined inter-rater agreement for abbreviated versions of the PSE-8 and PSE-9 when used by nonpsychiatric raters (Cooper, Copeland, Brown, Harris, & Gourlay, 1977; Wing, Nixon, Mann, & Leff, 1977; Rodgers & Mann, 1986). Cooper et al. (1977) examined the agreement between ratings of a psychiatrist or psychologist and those obtained by a sociologist or sociology graduate student. Agreement was evaluated for both joint interviews and test-retest over one week. For joint interviews, with the exception of situational anxiety ($r = 0.34$), inter-rater agreement for the remaining 13 section scores was good, with correlations ranging from 0.65 to 0.96 (mean $r = 0.77$). Test-retest reliability was lower with five section scores having correlations below 0.40 and the mean for the 14 sections decreasing to 0.49. The correlation between total scores was 0.88 for inter-rater agreement and 0.71 for test-retest. Finally, presence vs. absence decisions for the 150 rated items indicated good reliability with a mean inter-rater kappa of 0.74, and a mean test-retest kappa of 0.54.

Wing et al. (1977) conducted two studies of a brief version of the PSE-9. In the first, 95 patients were interviewed independently (5–84 days between interviews) by a nonmedical interviewer and a psychiatrist. Agreement was examined for 13 syndromes and was unacceptably low with a mean kappa of 0.34 (range 0–0.49). The authors examined agreement on five symptoms relating to anxiety and depression. Poor agreement was found for these symptoms with kappas below 0.32. In the second study, 28 interviews were conducted by a nonmedical interviewer. Audiotapes of these interviews were rated by a psychiatrist. The mean kappa for syndrome scores was 0.52 (range = 0.25–0.85). Ratings of the five symptoms yielded kappas above 0.62 with the exception of free-floating anxiety ($K = 0.34$).

In a large population study, Rodgers and Mann (1986) assessed inter-rater agreement between nurses and a psychiatrist's rating of audiotapes. Audiotapes of 526 abbreviated PSE-9 interviews were evaluated. A statistic of index of association was used, although the authors report that this measure was highly correlated with kappa. Of 44 symptoms rated, six were considered too infrequent to evaluate. Of the remaining 38 symptoms the median index of association was 0.73 (range 0–0.96); seven items (18%) were unacceptably low in level of agreement (index of association less than 0.45): Expansive Mood, Ideomotor Pressure, Obsessional Checking/Repeating, Obsessional Ideas/Rumination, Hypochondriasis, Suicidal Plans or Acts, and Ideas of Reference. Thirteen

syndrome scores derived from symptom ratings ranged from 0.29–0.94 (median = 0.76). Two syndrome scores were unacceptably low in agreement, Ideas of Reference (0.44) and Hypochondriasis (0.29).

In a recent study Okasha, Sadek, Al-Haddad and Abdel-Mawgoud (1993) examined rater agreement for assigning diagnosis based on *ICD-9*, *ICD-10*, and *DSM-III-R* criteria. The Arabic version of the PSE-9 was modified to collect extra data needed to make *ICD* and *DSM-III-R* diagnoses. One hundred adult inpatients and outpatients were interviewed by a single rater. An abstract form with PSE scores and other demographic and clinical information was then rated and diagnoses assigned. Overall kappa for nine broad diagnostic categories was acceptable (*ICD-9*, $K = 0.79$; *ICD-10*, $K = 0.82$; *DSM-III-R*, $K = 0.64$). Overall kappa values for more specific 18 diagnoses diminished somewhat but remained adequate (*ICD-9*, 0.62; *ICD-10*, 0.80, *DSM-III-R*, 0.63). Although this study indicates that PSE-9-derived information can be used to assign *ICD* and *DSM* diagnoses reliably it does not address the reliability of PSE-9 interviews themselves as diagnostic ratings were made from a single PSE abstract.

4.05.2.3.2 Supplements to the PSE

Two supplements to the PSE have been developed to address limitations in this instrument. These supplements address the assessment of lifetime psychopathology (McGuffin, Katz, & Aldrich, 1986) and change ratings (Tress, Bellenis, Brownlow, Livingston, & Leff, 1987). Because of the PSE's focus on the last month, its use in epidemiological studies is somewhat limited as these population-based investigations generally require the assessment of lifetime psychopathology. This concern led McGuffin et al. (1986) to modify the PSE. A Past History Schedule was developed to determine the dates of onset of worst episode of psychopathology, first psychiatric contact, and severest disturbance and recovery. Based on information obtained with the Past History Schedule, the PSE is then administered in three time formats: focusing on the last month, the most serious past episode, and modifying each PSE obligatory question with "have you ever experienced this?" Reliability assessment of this modified PSE using audiotaped interviews (McGuffin et al., 1986) has suggested adequate inter-rater agreement for the PSE CATEGO classes for past month (kappa range = 0.48–0.74), first episode (kappa range = 0.87–1), and ever (kappa range = 0.88–0.92). Rater agreement for dating past episodes was also found to be satisfactory

(rank-order correlation coefficients, median = 0.83, range 0.54–0.99).

Tress et al. (1987) modified the PSE for purposes of obtaining change ratings. The authors suggest that the advantage of the PSE over other instruments available for ratings of clinical change are that the PSE gives data for clinical classification, provides clear definitions of items, and uses a structured interview format. The PSE Change Rating is administered following a standard PSE assessment. Items not rated positively on the initial assessment are discarded (as well as items that were not rated). Subsequent ratings are only made on these remaining items. These items are subsequently rated on an eight-point scale from zero (Completely Remitted) to seven (Markedly Worsened). Inter-rater agreement based on observed interviews was high for grouped symptom ratings (ICC range = 0.75–0.99) and selected individual symptoms (ICC range = 0.70–1).

4.05.2.3.3 Summary

As the first semistructured clinical interview the PSE has an extensive history with application in a number of studies. Additionally, the PSE has been translated into over 40 languages and has been employed in cross-cultural studies. A potential advantage of the PSE is that it is not constrained by a particular diagnostic system; however, the PSE-10 was designed to yield *ICD-10* and *DSM-III-R* diagnoses (Wing et al., 1990). The reliability data for the PSE are constrained in that assessments have included a variety of versions and modifications of the PSE using raters with a variety of training with different populations. Caution should be exercised in applying these data to an investigator's own intended use. Furthermore, reliability data for the PSE-10, which has undergone substantial revision, are not yet available, although a multisite investigation has been conducted (Wing et al., 1990). Additionally, examination of diagnostic reliability achieved with the PSE, while encouraging, has been limited to a few diagnoses and are not available for *DSM-IV*.

4.05.2.4 Structured Clinical Interview for *DSM-IV*/Axis I Disorders

The Structured Clinical Interview for *DSM-IV* (SCID-I) is a semistructured interview designed to assist in determining *DSM-IV* Axis I diagnoses (First, Gibbon, Spitzer, & Williams, 1996). Construction of the interview began in 1983 following the introduction of the *DSM-III*, which introduced operationalized, specific be-

havioral criteria. At this time existing clinical structured diagnostic interviews became limited in that they did not conform to the *DSM-III* criteria (e.g., the SADS and PSE). Although the Diagnostic Interview Schedule (DIS) was developed to yield *DSM-III* diagnoses, the DIS was designed to be used by lay interviewers in epidemiological studies. It was argued by Spitzer (1983) that the most valid diagnostic assessment required the skills of a clinician so that the interviewer could rephrase questions, ask further questions for clarification, challenge inconsistencies, and use clinical judgment in ultimately assigning a diagnosis. Thus, the SCID was initially developed as a structured, yet flexible, clinical interview for *DSM-III*, and subsequently *DSM-III-R*, diagnoses (Spitzer, Williams, Gibbon, & First, 1992).

The SCID-I has been revised several times due to criteria changes and field trials. The interview was primarily developed for use with adults, but may be used with adolescents. It is contraindicated for those with less than an eighth grade education, severe cognitive impairments, and experiencing severe psychotic symptoms (First et al., 1996). The SCID-I is available in Spanish, German, Portuguese, Dutch, and Hebrew, as well as English.

Separate versions of the SCID-I have been developed for research and clinical applications. The clinical version, the SCID-I-CV, is briefer than the research version and focuses primarily on key diagnostic information (excluding the supplementary coverage provided in the research version) and on the most commonly occurring diagnoses (First et al., 1996). Within the research version, three variations of the interview provide differing comprehensive coverage of the disorders, subtypes, severity, course specifiers, and history. The research versions have been used historically for inclusion, exclusion, and data collection of study participants (in over 100 studies), and are distributed in loose page format to allow the investigator to customize the SCID-I to meet the needs of their research. The SCID-P (patient edition) was designed to address psychiatric patients and provides thorough coverage of psychotic disorders and past psychiatric history. The SCID-NP (nonpatient edition) was developed to focus on nonpsychiatric patients, and subsequently screens for psychotic disorders and provides less comprehensive coverage of psychiatric history. The SCID-P with Psychotic Screen was developed for patients where a psychotic disorder is not expected (and therefore only screens for psychotic disorders), but has thorough coverage of psychiatric history.

The SCID-I can usually be administered in 60 to 90 minutes, contingent on the quantity of

symptoms and disorders, and the ability of the interviewee to describe problems succinctly. It begins with an introductory overview followed by nine diagnostic modules. The overview provides open and closed questions that not only gather background information, but allows the interviewer to establish rapport with the interviewee before more detailed (and potentially uncomfortable) diagnostic questions are asked. The overview gathers information on demographics, work history, medical and psychiatric history, current stressors, substance use, and the interviewee's account of current and past problems (First et al., 1996).

There are nine diagnostic modules focusing on both current (usually defined as the past month) and lifetime assessment of diagnostic criteria: Mood Episodes, Psychotic Symptoms, Psychotic Disorders, Mood Disorders, Substance Use Disorders, Anxiety Disorders, Somatoform Disorders, Eating Disorders, and Adjustment Disorders. An optional module covers Acute Stress Disorder, Minor Depressive Disorder, Mixed Anxiety Depressive Disorder, and symptomatic details of past Major Depressive/Manic episodes. Each page of the modules contains questions, reprinted *DSM-IV* criteria, ratings, and instructions for continuation. Initial questions are closed-ended and followed up with open-ended elaboration questions. If further clarification is needed, the interviewer is encouraged to ask supplementary (their own) questions, give examples, present hypothetical situations, and challenge inconsistencies. In essence, the interviewer is testing diagnostic hypotheses. The ratings are based not on the question response, but on fulfillment of *DSM-IV* criteria which are provided alongside the questions. The interviewer is encouraged to use alternate sources of information to assist in rating the criteria, such as observed behavior, medical records, and informants. Each criteria is rated as one of the following: ? = inadequate information, 1 = symptom clearly absent or criteria not met, 2 = subthreshold condition that almost meets criteria, and, 3 = threshold for criteria met.

Unlike other diagnostic interviews such as the SADS, PSE, or DIS, where diagnostic algorithms are applied following the interview, the SCID incorporates diagnostic criteria and decision making within the interview. The use of a "decision-tree approach" allows the interviewer to test hypotheses and concentrate on more problematic areas (Spitzer, Williams, Gibbons & First, 1992). In addition, this approach makes the interview more time efficient, allowing the interviewer to "pass over" areas of no concern. Following the interview, the interviewer is provided with concise summary

scoring sheets to indicate the lifetime absence or threshold, and current presence of each disorder.

As a prerequisite for the SCID-I, the interviewer must possess adequate clinical experience and knowledge of psychopathology and diagnostic issues. The test developers recommend the following training: reading the administration sections of the *User's guide for the SCID-I*, reading the entire test booklet, reading the questions orally, practicing the SCID-I on a colleague/friend, watching a six-hour didactic training videotape titled *SCID-I 201*, role playing sample cases in the *User's guide for the SCID-I*, administering on an actual subject, conducting joint interviews (with independent ratings) followed by discussion sections, and examining inter-rater and test-retest reliability among interviewers (First et al., 1996). The following training materials and services are available: *User's guide for the SCID-I-I*, *SCID-I 201* video tape, videotape samples of interviews, on-site training, off-site monitoring, and SCID-I certification (under development). Following training, interviewers would benefit from ongoing supervision and feedback from an experienced SCID-I interviewer.

4.05.2.4.1 Reliability

Inter-rater agreement for the *DSM-III-R* version of the SCID-I was examined for 592 patients in five inpatient sites (one in Germany) and two nonpatient sites (Williams et al., 1992). At each site, two clinicians independently interviewed and diagnosed patients at least 24 hours but less than two weeks apart. In order to limit access to other information (e.g., chart review), interviewers were provided with only a brief summary of the hospital admission evaluation (circumstances of admission, number of prior hospitalizations, presenting problems). Diagnostic terms were excluded from the summary. For patients, overall weighted kappa was 0.61 for 18 current and 0.68 for 15 lifetime *DSM-III-R* diagnoses common to these sites. Disorders with poor agreement (i.e., *Ks* below 0.50) were current diagnoses of dysthymia, agoraphobia without panic disorder, and social phobia, and the lifetime diagnosis of agoraphobia without panic disorder. Agreement for specific substance dependence diagnoses at a drug and alcohol treatment facility was high with all diagnoses having kappas above 0.61 except cannabis dependence and polydrug dependence (both kappas below 0.35). For nonpatients, overall weighted kappa was 0.37 for five current diagnoses and 0.51 for seven lifetime diagnoses common to these sites. The only diagnoses in nonpatients with a kappa of 0.50 or greater were current panic disorder,

and lifetime diagnoses of alcohol dependence/abuse, other drug dependence/abuse, and panic disorder. Due to low occurrences, data were inconclusive for infrequent diagnoses.

Although generally satisfactory, these findings do indicate low agreement for some diagnoses. Williams et al. (1992) suggest several possible causes for low rater agreement in this study including the restriction of noninterview information, the focus on a broad range of diagnoses, and the flexible nature of the SCID in using clinical judgment. With regard to this last point, a review of a sample of audiotapes indicated that diagnostic disagreements were largely due to one interviewer's acceptance of a yes response without requesting elaboration while the other interviewer asked follow-up questions that ultimately determined that an initial yes response did not meet diagnostic criterion. As concluded by Williams et al. (1992) maximizing reliability on the SCID requires extensive training in the diagnostic criteria and an emphasis on not taking shortcuts but requiring that descriptions of behavior are elicited to justify each criterion.

Several other studies offer data on the reliability of the SCID-I, but the findings are confounded by small number of participants, changing *DSM* criteria and SCID-I revisions during the study, low base rates of disorders, and limited range of disorders (Segal, Hersen, & Van Hasselt, 1994). However, higher inter-rater agreement was observed in these studies ($K = 0.70-1$) compared to that obtained by Williams et al. (1992). The differences may have been due to the use of joint interviews (which controls for subject report variance) rather than independent interviews, access to noninterview information such as medical records and reports from other clinical staff, and the focus on a narrower range of diagnoses assessed.

4.05.2.4.2 Summary

The SCID is a well-established structured interview for determining *DSM-III-R* and *DSM-IV* diagnoses. Users may find the inclusion of diagnostic algorithms within the SCID and the use of skip-outs to result in a time-efficient interview. Reliability data from multiple sites indicate that the SCID can provide reliable *DSM-IV* diagnoses. Additionally, the SCID has some of the most extensive training materials and support available for any structured interview. The interview, user's guide, and all training materials have been completely updated for *DSM-IV*.

There are, however, a few disadvantages of the SCID-I. The interview does not cover a number of disorders, including infant, child-

hood, adolescence, cognitive, factitious, sexual, sleep, and impulse control disorders. Also, for those individuals interested in other diagnostic nosologies or needing to obtain broader clinical assessments, the restriction of the SCID to *DSM-IV* might be limiting. As with other structured interviews there is as yet no information currently available on the reliability of the SCID-I for the *DSM-IV* criteria. However, minor changes in the diagnostic criteria should not adversely affect reliability obtained with the *DSM-III-R* version.

4.05.2.5 Comprehensive Assessment of Symptoms and History

The Comprehensive Assessment of Symptoms and History (CASH; Andreasen, 1987) was developed without adherence to existing diagnostic systems (such as the *DSM* or *ICD*). The CASH adopted this approach based on observations that diagnostic criteria change over time and that methods of collecting information that conform to these criteria may be quickly outdated (Andreasen, Flaum, & Arndt, 1992). The CASH was designed for the study of psychosis and affective syndromes and is intended to provide a standardized assessment of psychopathology that will, ideally, yield diagnoses based on multiple criteria (both existing and future).

The CASH consists of nearly 1000 items divided into three sections: present state, past history, and lifetime history. The present state section consists of sociodemographic information intended to establish rapport and, subsequently, items pertaining to present illness. This section includes symptoms relating to the psychotic syndrome, manic syndrome, major depressive syndrome, treatment, cognitive assessment (laterality and a modified Mini-Mental Status Examination), a Global Assessment scale, and a summary of diagnoses for current episode.

Past history includes history of onset and hospitalization, past symptoms of psychosis, characterization of course, and past symptoms of affective disorder. To provide a detailed evaluation of phenomenology over time, for each symptom or sign, interviewers determine whether it was present during the first two years of illness, and whether it has been present for much of the time since onset. Finally, the lifetime history section includes history of somatic therapy, alcohol and drug use, pre-morbid adjustment, personality (schizotypal and affective characteristics), functioning in the past five years, Global Assessment scale and diagnoses for lifetime. Most items are given detailed definitions with suggested interview

probes. Items are typically rated on a six-point Likert-type scale.

A number of measures are embedded within the CASH. Scales within the CASH include the Scale for Assessment of Negative Symptoms (Andreasen, 1983), Scale for Assessment of Positive Symptoms (Andreasen, 1984), most items for the Hamilton depression scale (Hamilton, 1960), and the Brief Psychiatric Rating Scale (Overall & Gorham, 1962), the Mini-Mental Status Exam (Folstein, Folstein, & McHugh, 1975), and the Global Assessment Scale (Endicott, Spitzer, Fleiss, & Cohen, 1976). These measures make the CASH useful for repeat assessments.

The CASH was intended for use by individuals with experience and training in working with psychiatric patients (e.g., psychologists, psychiatrists, nurses, or social workers). A training program has been developed for its use, which includes training videotapes conducted with patients presenting a range of psychopathology. Narratives and calibrated ratings for the CASH items are available from the authors

4.05.2.5.1 Reliability

A small reliability study conducted with 30 patients has been reported (Andreasen et al., 1992). Two forms of rater agreement were evaluated. First, patients were interviewed by a primary rater with a second rater observing and asking clarifying questions when necessary. Second, test-retest reliability was evaluated with a third rater interviewing the patient within 24 hours of the initial interview. All raters had access to medical records. Agreement between the two initial raters was generally good for the spectrum diagnoses (Schizophrenia Spectrum, $K = 0.86$; Affective Spectrum, $K = 1$). For specific *DSM-III-R* diagnoses (focusing on diagnoses with more than one case), the results were positive for schizophrenia ($K = 0.61$), bipolar affective disorder ($K = 1$), and major depression ($K = 0.65$). However, reliability for schizoaffective disorder was somewhat low ($K = 0.45$). Test-retest reliability was similarly positive with kappas above 0.74 for spectrum diagnoses and above 0.64 for specific *DSM-III-R* diagnoses with the exception of schizoaffective disorder ($K = 0.52$).

Because of the intent of the CASH to provide a reliable assessment of symptoms and functioning independent of diagnostic classification, it is important to examine the reliability of the individual items. Given the number of items, Andreasen et al. (1992) provide summaries of the intraclass correlation coefficients for the inter-rater and test-retest administrations. For

inter-rater agreement, ICCs were generally high with three-quarters of the items having ICCs greater than or equal to 0.65. For the test-retest design, reliability was somewhat lower with approximately one-half of the items demonstrating ICC greater than or equal to 0.65.

Reliability data have been published for some more critical items or content areas (Andreasen et al., 1992). For history of illness, ICC values for both inter-rater and test-retest designs were quite adequate with values generally above 0.60 (median ICCs above 0.70). Reliability for items relating to manic and depressive syndromes was acceptable (median ICCs = 0.68 and 0.58, respectively). For positive and negative symptoms inter-rater and test-retest reliability was generally acceptable for "current" and "much of time since onset" time frames (global symptom scores ICCs greater than 0.65). However, test-retest reliability for negative symptoms rated for the "first two years of illness" and "worst ever" were unacceptably low (ICCs = 0 and 0.48, respectively). Test-retest data on premorbid and prodromal symptoms were very low (median ICCs = 0.37 and 0.25, respectively), while residual symptom ratings were somewhat better (median ICC = 0.60).

4.05.2.5.2 Summary

The CASH presents several advantages including its lack of adherence to any diagnostic system. This may afford the opportunity to collect a rich body of information on individuals. The comprehensiveness of the items is intended to allow for diagnoses for *DSM* and *ICD* to be generated while not narrowing the collection of information to these systems. Available reliability data are encouraging, with some exceptions as noted above. The availability of training materials including videotapes and consensus ratings is also attractive. The CASH also has companion instruments that are useful in the context of longitudinal assessments, providing baseline and follow-up assessment of psychosocial functioning and symptomatology.

The CASH is limited in several respects. First, because it seeks a full assessment of symptoms and history without regard to diagnostic criteria the entire CASH must be administered (however, some syndromes can be skipped if the interviewer already knows that a syndrome is not applicable). With nearly 1000 items this ensures lengthy assessment. Second, the CASH is limited to schizophrenia and affective syndromes and alcohol and drug use. Thus, it may not provide the breadth of symptom and diagnostic evaluation that some settings may require. Finally, although intended to be

capable of assigning diagnoses based on extant nosological systems, the CASH may not always be capable of achieving this goal. Diagnostic criteria may require symptom information that does not conform to the information obtained with the CASH. Interested users should carefully evaluate the content of the CASH to ensure that relevant diagnoses can be made.

4.05.2.6 Diagnostic Interview for Genetic Studies

The Diagnostic Interview for Genetic Studies (DIGS; Nurnberger et al., 1994) was developed by participants in the NIMH Genetics Initiative. The need for the DIGS arose from perceptions that inconsistent findings in the genetics of psychiatric illnesses might be, in part, the result of differences in phenotypic assessment. Problems in assessment are highlighted in genetics research where family members are more likely to evince spectrum disorders and subclinical symptomatology.

The DIGS adopts a polydiagnostic approach that collects clinical information in sufficient detail to allow a variety of diagnostic definitions to be applied including *DSM-III-R* (a new version for *DSM-IV* is now available), modified RDC, RDC, Feighner criteria, *ICD-10*, and the European Operational Criteria (OPCRIT) Checklist (McGuffin & Farmer, 1990). As with the CASH, the advantage of this feature includes the collection of a broad data set for diagnostic entities whose definitions are sometimes ambiguous and evolving. However, the DIGS (unlike the CASH) explicitly collects information that conforms to several diagnostic systems including *DSM-IV*. Items from other interviews have been incorporated into the DIGS including the SADS, CASH, and DIS.

Like the SADS the DIGS provides standard probe questions and criterion-based definitions. Additionally, the DIGS requires clinical judgment for item ratings and in determining the need for further probe questions. Sections of the DIGS begin with one or two screening questions that, if denied, allow the interviewer to skip out of the remainder of the section. Questions are integrated so as to cover the various diagnostic criteria covered while maintaining an efficient flow of questioning. The interview can take 30 minutes to four hours depending on the complexity of the symptomatology (median time for symptomatic individuals is two and one-half hours).

The DIGS begins with a modified Mini-Mental Status examination in order to determine if the interview should be terminated as a result of cognitive impairment. The Introduction continues with demographics and extensive

medical history screening questions. Somatization follows to enhance flow from medical history. An overview section assesses psychiatric history and course of illness and this information is summarized graphically in a time line to provide chronology of symptoms and episodes of illness. Mood disorders include major depression, mania/hypomania, and cyclothymic personality disorder. The DIGS also provides a detailed assessment of substance use history. Psychotic symptoms are assessed in great detail and psychotic syndromes are distinguished. Additionally, schizotypy is also assessed. A unique feature of the DIGS is an assessment for comorbid substance use. The aim of this section is to determine the temporal relationship between affective disorder, psychosis, and substance use. Suicidal behaviors, major anxiety disorders (except generalized anxiety disorder), eating disorders, and sociopathy are also evaluated. Finally, at the conclusion of the DIGS, the interviewer completes a Global Assessment scale (Endicott et al., 1976), the Scale for the Assessment of Negative Symptoms (Andreasen, 1983), and the Scale for the Assessment of Postive Symptoms (Andreasen, 1984).

Appropriate personnel to administer the DIGS are mental health professionals with clinical experience and familiarity with multiple diagnostic systems. In the study of Nurnberger et al. (1994), all but one interviewer had prior experience administering semistructured clinical interviews. Training as outlined in the reliability studies (Nurnberg et al., 1994) consisted of demonstration interviews by senior clinicians, the administration of a DIGS under supervision, and supplemental training involving three videotaped patient interviews.

4.05.2.6.1 Reliability

Test-retest reliability for the DIGS has been evaluated for major depression, bipolar disorder, schizophrenia, schizoaffective disorder and an "other" category (Faraone et al., 1996; Nurnberger et al., 1994). Test-retest reliability was evaluated within participating research sites as well as across sites. For the intrasite study, participants were independently interviewed with the DIGS over a period of no more than three weeks. For the intersite study, interviewers traveled to other research centers so that interviewers from different sites could assess the same subjects. These pairs of interviews were conducted within a 10-day period. With the exception of *DSM-III-R* schizoaffective disorder (K less than 0.50), *DSM-III-R* and RDC target diagnoses showed excellent reliability with kappas above 0.72 across the two studies.

4.05.2.6.2 Summary

The DIGS appears to be an excellent instrument for the study of the genetics of schizophrenia and affective disorders and other comorbid conditions. It provides an exhaustive assessment of symptomatology that allows for the comparison of findings across a number of diagnostic systems. Furthermore, it targets spectrum disorders and other comorbid conditions that may be relevant in family studies of schizophrenia and affective disorders. Although reliability has been shown to be high across different sites, these data are limited to schizophrenia and the affective disorders (but are not available for bipolar II and schizotypal personality); data are not available for other disorders such as the anxiety and eating disorders. As emphasized by the developers, the DIGS is not designed for routine clinical use. It is designed to be used by highly trained clinical interviewers for use in research settings.

4.05.2.7 Diagnostic Interview Schedule

The DIS (Robins, Helzer, Croughan, & Ratcliff, 1981) is a highly structured interview developed for use in large-scale epidemiological studies by the NIMH (the Epidemiological Catchment Area, ECA, projects). Because of the logistical constraints in general population studies, the use of traditional structured interviews administered by clinicians is prohibitive. The DIS was developed for use by lay interviewers with relatively brief training (one to two weeks). Thus, unlike structured interviews such as the SADS or SCID, the DIS minimizes the amount of discretion that an interviewer exercises in either the wording of questions or in determining the use of probe questions. Additionally, diagnoses are not made by the interviewer; rather, the data are scored, and diagnoses assigned, by a computer program.

The DIS was designed to provide information that would allow diagnoses to be made according to three diagnostic systems: *DSM-III* (APA, 1980), the Feighner criteria (Feighner et al. 1972), and the RDC. The interview covers each item or criterion in the form of one or more close-ended questions. Questions assess the presence of symptoms, whether they meet criteria for frequency and clustering in time, and whether they meet the age-at-onset criterion. The use of a Probe Flow Chart provides probes needed to determine severity and address alternative explanations. Rules concerning when and what probe questions to use are explicit in the interview. Nearly all questions can be read by lay interviewers as written. The DIS can take between 45 and 75 minutes to complete.

4.05.2.7.1 Reliability

vRobins et al. (1981) addressed the question of whether lay interviewers could obtain psychiatric diagnoses comparable to those obtained by a psychiatrist. In a test-retest design, subjects (mostly current or former psychiatric patients) were separately interviewed by a lay interviewer and a psychiatrist, both using the DIS. With the exception of *DSM-III* panic disorder ($K = 0.40$), kappas for all lifetime diagnoses across each diagnostic system were 0.50 or greater. Mean kappas for each diagnostic system were quite adequate: *DSM-III*, $K = 0.69$; Feighner, $K = 0.70$; RDC, $K = 0.62$. Further analysis of these data (Robins, Helzer, Ratcliff, & Seyfried, 1982) suggested that current disorders and severe disorders are more reliably diagnosed with the DIS than disorders in remission or borderline conditions.

Although the findings of Robins et al. (1981, 1982) suggested acceptable concordance between lay interviewers and psychiatrists using the DIS, these data do not address whether the DIS would yield similar diagnoses as obtained by psychiatrists with broader clinical assessment than that allowed by the DIS alone. Additionally this first study was conducted with a largely psychiatric sample and may not be generalizable to the nonclinical populations the DIS was designed for. Anthony et al. (1985), using data from the ECA obtained in Eastern Baltimore, compared DIS-obtained diagnoses with psychiatrist-conducted clinical reappraisal examinations ($N = 810$). These clinical reappraisals were based on an augmented PSE consisting of 450 signs and symptoms and included all items of the PSE-9. Additionally, psychiatrists reviewed all available records. The two assessments were independent and the majority were completed within four weeks of each other. Diagnostic comparisons were for conditions present at the time of the interview or within one month prior to the interview. Results indicated substantial differences between DIS and psychiatrists' diagnoses. Except for schizophrenia and manic episode, one month prevalence rates for *DSM-III* diagnostic categories were significantly different for the two methods. Additionally, there was very low concordance between DIS-based diagnoses and those obtained by psychiatrists with kappas for selected diagnoses all below 0.36.

In a second major study based on ECA data collected in St. Louis, Helzer et al. (1985) compared lay-interview diagnoses with those obtained by a psychiatrist ($N = 370$). The psychiatrist re-examined subjects with the

DIS and were also allowed to ask whatever supplemental questions they deemed necessary to resolve diagnostic uncertainties following the DIS. The majority of psychiatrist interviews were conducted within three months of the lay interview. Diagnostic comparisons were made for lifetime diagnoses. Helzer et al.'s summary of the data was somewhat optimistic, indicating that corrected concordance was 0.60 or better for eight of the 11 lifetime diagnoses. However, when kappa statistics are examined, the results are less promising. Only one of eleven diagnoses obtained a kappa greater than or equal to 0.60, and eight diagnoses had kappas below 0.50. As summarized by Shrout et al. (1987) "for most diagnoses studied, the agreement in community samples between the DIS and clinical diagnoses is poor" (p. 177).

A number of other studies have been conducted comparing lay interviewer-administered DIS diagnoses with clinical diagnoses (e.g., Erdman et al., 1987; Escobar, Randolph, Asamen, & Karno, 1986; Ford, Hillard, Giesler, Lassen, & Thomas, 1989; Spengler & Wittchen, 1988; Wittchen, Semler, & von Zerssen, 1985; also see recent review by Wittchen, 1994). These studies are difficult to summarize and their interpretability is sometimes limited due to the use of a variety of assessment methodologies, diagnostic systems, and populations. Although some diagnoses achieve acceptable concordance levels between the DIS and clinical diagnoses, in total the results of these studies suggest limitations of the DIS consistent with the observations of Shrout et al. (1987). Wittchen (1994) has summarized particular problems apparent in the panic, somatoform, and psychotic sections of the DIS.

4.05.2.7.2 Summary

The DIS marked a significant development in the epidemiological study of psychopathology. The ECA findings based on the DIS have yielded important information about the epidemiology of a variety of disorders. However, studies examining the concordance between the DIS and clinical interviews conducted by psychiatrists suggest that there may be appreciable diagnostic inaccuracy in DIS-based diagnoses. Although the use of the DIS in epidemiological studies may continue to be warranted given the logistical constraints of such studies and the important data the DIS does obtain, the concern with diagnostic reliability should preclude the use of the DIS in settings where other structured diagnostic interviews can be used (e.g., the SADS or SCID).

4.05.2.8 Composite International Diagnostic Interview

The Composite International Diagnostic Interview (CIDI; Robins et al., 1988; World Health Organization [WHO], 1990) was developed at the request of WHO and the United States Alcohol, Drug Abuse, and Mental Health Administration. The CIDI was designed to serve as a diagnostic instrument in cross-cultural epidemiological and comparative studies of psychopathology. The initial version of the CIDI was based on the DIS, to cover DSM-III diagnoses, and initially incorporated aspects of the PSE since the PSE has been used in cross-cultural studies and reflects European diagnostic traditions. Some items from the DIS were altered either to provide further information for the PSE or to address language and content that would allow cross-cultural use. Additional questions were added to provide adequate coverage of the PSE-9. These PSE items were rewritten into the closed-ended format of the DIS.

Initial versions of the CIDI provided *DSM-III* diagnoses and updated versions, used in Phase II WHO field trials, now provide *DSM-III-R* and *ICD-10* diagnoses (WHO, 1990). The latest version of the CIDI has also eliminated all questions that are not needed for *DSM-III-R* (deletion of Feighner and RDC criteria) and has added items to assess *ICD-10* criteria. Furthermore, the PSE approach was abandoned and only same questions from the PSE were retained because they were relevant to *ICD-10* diagnoses. Revisions of the CIDI to meet *DSM-IV* criteria are in progress. A Substance Abuse Module was developed for the CIDI to be used alone or substituted for the less detailed coverage of drug abuse in the CIDI proper (Cottler, 1991). Other modules that have been developed or are under development include post-traumatic stress disorders, antisocial disorder, conduct disorder, pathological gambling, psychosexual dysfunctions, neurasthenia, and persistent pain disorder (Wittchen, 1994).

Like the DIS, the CIDI was intended to be used by lay interviewers with modest training (one week), and to be capable of rapid administration. In a multicenter study conducted in 15 countries the CIDI was found to be judged appropriate by the majority of interviewers (Wittchen et al., 1991). However, 31% of interviewers rated parts of the CIDI as inappropriate, in particular sections for schizophrenia and depression. The CIDI also takes a long time to administer: one-third of the interviews took one to two hours and another third lasted two to three hours (Wittchen et al., 1991). In this study the duration of the

interviews may have been extended because of the assessment of predominantly clinical populations. One might expect briefer administration times with general population samples.

As with the DIS, training for the CIDI can be conducted in five days. No professional experience is necessary as the CIDI is intended to be used by lay interviewers. Training sites participating in the WHO field trials may be available to conduct courses. However, there is a CIDI user manual, a standardized training manual with item-by-item specifications, and computer scoring program available from WHO (1990).

4.05.2.8.1 Reliability

In an evaluation of the "prefinal" version of the CIDI (*DSM-III* and PSE) across 18 clinical centers in 15 countries, Wittchen et al. (1991) found high inter-rater agreement. Kappas for diagnoses were all 0.80 or greater with the exception of somatization (0.67). Wittchen (1994) has summarized test-retest reliability of the CIDI across three studies involving independent interviews conducted within a period of three days. Kappa coefficients for *DSM-III-R* diagnoses were all above 0.60 with the exception of bipolar II (0.59), dysthymia (0.52), and simple phobia (0.59).

Two studies have also examined the concordance between the CIDI and clinical ratings. Farmer, Katz, McGuffin, and Bebbington (1987) evaluated the test-retest concordance between CIDI PSE-equivalent items obtained by a lay interviewer and PSE interviews conducted by a psychiatrist. Interviews were conducted no more than one week apart. Concordance at the item level was unacceptably low. Of 45 PSE items, 37 (82%) achieved kappas below 0.50. Agreement was somewhat higher at the syndrome level but remained low for the syndromes of behavior, speech, and other syndromes (Spearman $r = 0.44$) and specific neurotic syndromes (Spearman $r = 0.35$).

Janca, Robins, Cottler, and Early (1992) examined diagnostic concordance between a clinical interviewer using the CIDI and a psychiatrist in a small sample of patients and nonclinical subjects ($N = 32$). Psychiatrists asked free-form questions and completed an *ICD-10* checklist following either the observation of lay-administered CIDI interview, or following their own administration of the CIDI. Overall diagnostic agreement appeared adequate with an overall kappa of 0.77. High concordance was also found for the *ICD-10* categories of anxiety/phobic disorders ($K = 0.73$), depressive disorders ($K = 0.78$), and psychoactive substance use disorders ($K = 0.83$).

4.05.2.8.2 Summary

The CIDI may be considered the next step beyond the DIS. This instrument incorporated lessons learned from the development of the DIS and has been subject to repeated revisions to enhance its reliability and cross-cultural application. The CIDI appears to have achieved somewhat better reliability than the DIS and it covers the latest diagnostic standards of *ICD-10* and *DSM-III-R* (soon to cover *DSM-IV*). However, the concordance between CIDI-obtained diagnoses and diagnoses obtained by other structured interviews administered by clinicians (e.g., SADS, SCID) remains unclear.

4.05.3 PERSONALITY DISORDERS

4.05.3.1 Structured Interview for *DSM-IV* Personality Disorders

Introduced in 1983, the Structured Interview for *DSM-III* Personality Disorders (SIDP) was the first structured interview to diagnose the full range of personality disorders in *DSM-III* (Stangl, Pfohl, Zimmerman, Bowers, & Corenthal, 1985). Subsequent versions have addressed Axis II criteria for the *DSM-III-R* (SIDP-R; Pfohl, Blum, Zimmerman, & Stangl, 1989) and the *DSM-IV* (SIDP-IV; Pfohl, Blum, & Zimmerman, 1995).

SIDP-IV is organized into 10 topical sections: Interests and Activities, Work Style, Close Relationships, Social Relationships, Emotions, Observational Criteria, Self-perception, Perception of Others, Stress and Anger, and Social Conformity. This format is intended to provide a more conversational flow and is thought to enhance the collection of information from related questions and facilitate the subsequent scoring of related criteria.

The SIDP-IV can be administered to a patient and to an informant, requiring 60–90 and 20–30 minutes, respectively (Pfohl et al., 1995). In addition, the interview takes 20–30 minutes to score. Each page of the interview provides questions, prompts, diagnostic criteria, and scoring anchors. The informant interview is composed of a subset of questions from the patient interview. Two alternate versions of the SIDP-IV are available. The SIDP-IV Modular Edition is organized by personality disorders rather than by topical sections. This modular form permits the interviewer to focus on disorders of interest and to omit others. The Super SIDP is an expanded version that includes all questions and criteria necessary to assess *DSM-III-R*, *DSM-IV*, and *ICD-10* personality disorders.

Several instructions for administering the SIDP-IV are noteworthy. First, the SIDP uses a

“five year rule” to operationalize criteria involving an enduring pattern that represents personality. Thus, “behavior, cognitions, and feelings that have predominated for most of the last five years are considered to be representative of the individual’s long-term personality functioning” (Pfohl et al., 1995, p. ii). The SIDP-IV is intended to follow assessment of episodic (Axis I) disorders in order to assist the interviewer in ruling out the influence of temporary states of behavior described by the patient. Second, the patient’s responses are not given a final rating until following the interview. This is intended to allow for all sources of information to be reviewed before rating. However, unlike previous versions of the SIDP, the SIDP-IV now provides a rater with the opportunity to rate or refer to specific *DSM-IV* criteria associated with each set of questions. Third, use of an informant is optional and Pfohl et al. (1995) note that while the frequency of personality diagnoses may increase with the use of informants there appears to be little effect on predictive validity.

Each diagnostic criterion is scored as one of the following: 0 = not present or limited to rare isolated examples, 1 = subthreshold: some evidence of the trait but it is not sufficiently pervasive or severe to consider the criterion present, 2 = present: criterion is present for most of the last five years (i.e., present at least 50% of the time during the last five years), and 3 = strongly present: criterion is associated with subjective distress or some impairment in social or occupational functioning or intimate relationships. Unlike other personality interviews (e.g., the IPDE and SCID-II), scores of both 2 and 3 count towards meeting diagnostic criteria (Pilkonis et al., 1995).

The SIDP-IV is an interview requiring clinical skill in determining the need for additional probe questions and to discriminate between personality (Axis II) disorders and episodic (Axis I) disorders. The developers of the SIDP-IV recommend one month of intensive training to administer the interview properly (Pfohl et al., 1995; Standage, 1989). Pfohl et al. (1995) have reported success with interviewers having at least an undergraduate degree in the social sciences and six months of previous experience with psychiatric interviewing. Training videotapes and workshops are available from the developers of the SIDP-IV.

4.05.3.1.1 Reliability

Several investigations of inter-rater reliability reveal poor to good agreement. Using the SIDP-R, Pilkonis et al. (1995) found that inter-rater agreement for continuous scores on either the

total SIDP-R score or scores from Clusters A, B, and C, was satisfactory (ICCs ranging from 0.82 to 0.90). Inter-rater reliability for presence or absence of any personality disorder with the SIDP-R was moderate with a kappa of 0.53. Due to infrequent diagnoses, mixed diagnoses, and the number of subthreshold protocols, in this study kappas for individual diagnoses were not provided.

Stangl et al. (1985) conducted SIDP interviews on 63 patients (43 interviews were conducted jointly, and 20 interviews were separated by up to one week). The kappa for presence or absence of any personality disorder was 0.66. Only five personality disorders occurred with enough frequency to allow kappa to be calculated: dependent (0.90), borderline (0.85), histrionic (0.75), schizotypal (0.62), and avoidant (0.45). Using the SIDP among a small sample of inpatients, Jackson, Gazis, Rudd, & Edwards (1991) found inter-rater agreement to be adequate to poor for the five specific personality disorders assessed: borderline ($K = 0.77$), histrionic (0.70), schizotypal (0.67), paranoid (0.61), and dependent (0.42).

The impact of informant interviews on the diagnosis of personality disorders and inter-rater agreement for the SIDP was assessed by Zimmerman, Pfohl, Stangl, and Corenthal (1986). Inter-rater agreement (kappa) for the presence or absence of any personality disorder was 0.74 before the informant interview and 0.72 after the informant interview. Kappas for individual personality disorders were all 0.50 or above. Reliability did not appear to benefit or be compromised by the use of informants. However, the informant generally provided additional information on pathology and, following the informant interview, diagnoses that had been established with the subject only were changed in 20% of the cases (Zimmerman et al., 1986).

In an examination of the long-term test-retest reliability of the SIDP, Pfohl, Black, Noyes Coryell, and Barrash (1990) administered the SIDP to a small sample of depressed inpatients during hospitalization and again 6–12 months later. Information from informants was used in addition to patient interviews. Of the six disorders diagnosed three had unacceptably low kappas (below 0.50): passive-aggressive (0.16), schizotypal (0.22), and histrionic (0.46). Adequate test-retest reliability was obtained for borderline (0.58), paranoid (0.64), and anti-social (0.84).

4.05.3.1.2 Summary

The SIDP-IV represents the latest version of the first interview to assess the spectrum of

DSM personality disorders. Although originally developed to be administered in a topical format to assess *DSM* personality disorders the SIDP-IV now provides alternative formats for assessing specific disorders without administering the entire SIDP-IV and for assessing *ICD* diagnoses. Reliability data are encouraging for some disorders. However, these data are limited to selected disorders using the SIDP and reliability data have not been presented for specific disorders using the SIDP-R (Pilkonis et al., 1995). No reliability data are available for the SIDP-IV. Little data are available concerning the long-term test-retest reliability of the SIDP. The SIDP-IV does not come with a screening questionnaire to assist in identifying personality disorders that might be a focus of the interview.

4.05.3.2 International Personality Disorder Examination

The International Personality Disorder Examination (IPDE; Loranger et al., 1995), a modified version of the Personality Disorder Examination (PDE), is a semistructured interview designed to assess personality disorders in both the *DSM-IV* and *ICD-10* classification systems. The PDE was initially developed in the early 1980s to assist in the diagnosis of personality disorders. At that time, only structured interviews existed that focused on Axis I mental disorders. A highly structured lay-administered interview for personality disorders was thought to be inappropriate due to the complexity of diagnostic criteria and level of inference required (Loranger et al., 1994). The first version of the PDE was completed in 1985. Beginning in 1985, international members of the psychiatric community attended several workshops to formulate an international version of the PDE, the IPDE, which was developed under the WHO, and the US Alcohol, Drug Abuse, and Mental Health Administration System. The purpose of the IPDE was to assess personality disorders in different languages and cultures.

The IPDE interview surveys behavior and life experiences relevant to 157 criteria and can be used to determine *DSM-IV* and *ICD-10* categorical diagnoses and dimensional scores of personality disorders. The IPDE is not recommended for use on individuals who are under the age of 18, agitated, psychotic, severely depressed, below normal intelligence, or severely cognitively impaired. The interview is available in the following languages: English, Spanish, French, German, Italian, Dutch, Danish, Norwegian, Estonian, Greek, Russian, Japanese, Hindi, Kannada, Swahili, and Tamil (Loranger et al., 1995).

The IPDE contains materials for determining both *DSM-IV* and *ICD-10* diagnoses. However, due to the long length of the interviews noted in the field trials (mean interview length was 2 hours, 20 minutes), the interview is distributed in two modules for each classification system. Furthermore, clinicians and researchers can easily administer specific personality modules to suit their purpose (Loranger et al., 1994).

A self-administered IPDE screening questionnaire may be administered prior to the interview in order to eliminate subjects who are unlikely to have any or a particular personality disorder. Similar to the SCID-II (described below), a corresponding low diagnostic threshold (for endorsement) is set for each question. If three or more items are endorsed for a specific personality disorder, then the interview is administered for that personality disorder.

In the attempt to establish reliable diagnoses, the IPDE interview utilizes two distinct guidelines. First, the behavior or trait must be present for at least five years to be considered an expression of personality with some manifestations (based on the disorder) occurring within the past 12 months (Loranger et al., 1995). This strict criterion is adopted to ensure the enduring nature of behavior and life experiences, and rule out transient or situational behaviors. A second guideline for the IPDE is that one criterion must be met before the age of 25. However, if an individual develops a personality disorder later in life (with no criterion exhibited prior to age 25) the IPDE provides an optional "late onset" diagnosis (Loranger et al., 1995).

The developers constructed the IPDE interview not only to be clearly organized, but to "flow" naturally. As a result, the diagnostic criteria are not ordered by cluster or disorder, but by six sections that assess major life domains: Work, Self, Interpersonal Relationships, Affects, Reality Testing, and Impulse Control. Each section begins with open-ended questions that provide a transition between sections and allow the interviewer to gather general background information. Closed-ended and elaboration questions follow for each criterion (Loranger et al., 1995). Each individual page of the IPDE is designed to optimally assist the interviewer in correctly determining if the diagnostic criteria is met. Each page contains: personality disorder and criterion number, structured questions, reprinted *DSM-IV* or *ICD-10* criteria, notes on determining criteria, descriptions of scoring criteria, and scoring areas for both interview and informants.

The scoring of the IPDE interview is similar to other semistructured interviews. Prior to the interview, the developers recommend that collecting information or conducting interviews

on Axis I disorders be completed to assist in scoring the criteria. Each trait or behavior (i.e., criterion) is scored as one of the following: absent or normal (0), exaggerated or accentuated (1), criterion level or pathological (2), and interviewee is unable to furnish adequate information (?). Some items may not be applicable to all interviewees and scored not applicable. The IPDE also allows the interviewer to rate each criterion based on informants (Loranger et al., 1995). The IPDE manual provides information on the scope and meaning of each criterion, and provides guidelines and anchor points for scoring. The manual does not recommend challenging the interviewee on inconsistencies with informants during the interview, due to the potential threat to rapport. However, challenging discrepancies occurring within the interview is encouraged. The IPDE may be hand scored or computer scored (program provided by publishers). Hand-scored algorithms and summary sheets are provided to assist in determining categorical diagnoses and dimensional scores.

The IPDE developers recommend that only those with the clinical experience and training to make psychiatric diagnoses independently use the IPDE (Loranger et al., 1994). The IPDE manual strongly discourages the use of IPDE by clinicians early in their training, research assistants, nurses, medical students, and graduate students. In addition, the interviewer should have familiarity with the *DSM-IV* and *ICD-10* classification systems (Loranger et al., 1995). The test manual recommends the following training: read the interview and manual thoroughly, practice on several participants to get familiar with the interview, administer with an IPDE-experienced interviewer, and discuss any problems in administration or scoring. Before administering the IPDE, the interviewer should have thorough knowledge of the scope and meaning of each criterion and scoring guidelines. IPDE training courses are offered at the WHO training centers.

4.05.3.2.1 Reliability

The IPDE field trial conducted in 11 countries and 14 centers evaluated inter-rater reliability in joint interviews as well as test-retest reliability with an average test-retest interval of six months (the test-retest interviews were conducted by the same interviewer). Results indicated overall weighted kappa for individual definite personality disorders to be 0.57 for the *DSM-III-R* and 0.65 for the *ICD-10* (Loranger et al., 1994). Median kappas for definite or probable personality diagnoses were 0.73 for *DSM-III-R* and 0.77 for *ICD-10*. Using

broader definite or probable criteria, kappa for any personality disorder increased to 0.70 for *DSM-III-R* and 0.71 for *ICD-10*. For temporal stability, kappas for the presence or absence of a personality disorder were 0.62 for *DSM-III-R* and 0.59 for *ICD-10*. Inter-rater reliability was higher for dimensional scores with ICCs ranging from 0.79 to 0.94 for the *DSM-III-R* and 0.86 to 0.93 for the *ICD-10*. Temporal stability for dimensional scores was also high with ICCs ranging from 0.68 to 0.92 for *DSM-III-R* and from 0.65 to 0.86 for *ICD-10*.

Pilkonis et al. (1995) also evaluated the reliability of the third version of the PDE. Intraclass correlations for total scores or cluster scores ranged from 0.85 to 0.92. Inter-rater agreement (kappa) for the presence or absence of any personality disorder was 0.55.

Loranger et al. (1991) examined inter-rater agreement and test-retest reliability of the PDE in a sample of psychiatric inpatients. Second administrations of the PDE were conducted one week to six months later by a separate interviewer blind to the initial assessment. Inter-rater agreement between two raters was assessed at both the initial and repeated interview. At the first interview kappas for inter-rater reliability ranged from 0.81 to 0.92 (median = 0.87). At the repeat interview kappas for inter-rater reliability ranged from 0.70 to 0.95 (median = 0.88). At follow up there was a significant reduction in the number of criteria met on all disorders except schizoid and antisocial. Stability of the presence or absence of any personality disorder was moderated with a kappa of 0.55.

O'Boyle and Self (1990) interviewed 20 patients with a depressive disorder for a personality disorder. Eighteen patients were re-interviewed across a mean of 63 days for the presence or absence of personality disorder. Intraclass correlations were 0.89 to 1 and inter-rater agreement (kappa) was 0.63. Depressive disorders did not consistently affect categorical diagnoses, but dimensional scores were higher during depressed periods.

4.05.3.2.2 Summary

The IPDE has a number of strengths to recommend its use. First, the IPDE (and PDE) has demonstrated medium to high inter-rater agreement and temporal reliability for both categorical diagnoses and dimensional scores. In addition, preliminary investigations into the influence of Axis I disorders (e.g. depression) on the assessment of personality disorders indicate no significant influence on PDE categorical diagnoses. Second, a detailed training manual accompanies the interview, which provides

thorough instructions and scoring algorithms. Third, a unique feature of the IPDE is dual coverage of the *DSM-IV* and *ICD-10* criteria. Fourth, in addition to providing categorical diagnoses, the IPDE measures dimensional scores which provide information about accentuated normal traits below the threshold required for a personality disorder. Finally, the IPDE is available in several languages and has been studied in 11 countries.

The IPDE, while quite comprehensive, is flexible enough to permit more economical administration. The *DSM-IV* and *ICD-10* modules can be administered separately. Furthermore, rather than administer in thematic organization, the IPDE can be limited to diagnostic modules of interest. A self-administered screening questionnaire is available to assist in identifying personality disorders that might be of focus in the interview.

Inter-rater agreement between the SCID-II and IPDE have led some researchers to conclude that the IPDE (and PDE) presents more stringent guidelines to fulfill personality disorder criteria (Hyler, Skodol, Kellman, Oldham, & Rosnik, 1990; Hyler, Skodol, Oldham, Kellham, & Doldge, 1992; O'Boyle & Self, 1990). This stringent determination is probably due to the consistent five-year time period requirement for personality traits. In conclusion, the specificity of the instrument is increased (fewer false positives) but this may be at the cost of decreased sensitivity (more false negatives).

4.05.3.3 Structured Clinical Interview for *DSM-IV* Personality Disorders

The Structured Clinical Interview for *DSM-IV* Personality Disorders (SCID-II) is a structured interview that attempts to provide an assessment of the 11 *DSM-III-R* personality disorders, including the diagnosis of self-defeating personality disorder, which is included in Appendix A of *DSM-III-R* (First, Spitzer, Gibbon, & Williams, 1995). The SCID-II interview can be used to make categorical or dimensional personality assessments (based on the number of items judged present). The SCID-II was developed as a supplementary module to the SCID-I, but was redesigned in 1985 to be a separate and autonomous instrument due to different assessment procedures and length of interview (First et al., 1995).

In conducting the SCID-II, it is extremely important to evaluate the interviewee's behavior out of the context of an Axis I disorder (Spitzer et al., 1990). The test developers recommend an evaluation of Axis I disorders prior to the

SCID-II with either a SCID-I or some other Axis I evaluation.

A self-report screening questionnaire is provided to improve time efficiency. Each of the 113 items on the questionnaire corresponds to a diagnostic criterion and is purposefully constructed to have a low threshold for a positive response, and is therefore for screening purposes only. Interviewers should probe all items coded "yes" in the questionnaire. Under most circumstances, the interviewer does not need to interview for the negatively endorsed criteria, due to the low probability of psychopathology. Negative questionnaire responses should be followed up when either the interviewer suspects that the criterion or personality disorder is actually present or when the number of items endorsed positively in the interview is only one item below that required for making a diagnosis (in which case all questions for that diagnosis should be probed).

Utilizing the screening questionnaire, the SCID-II interview can usually be administered in 30–45 minutes (First et al., 1995). First, Spitzer, Gibbons, Williams, Davies, et al. (1995) interviewed 103 psychiatric patients and 181 nonpatients, and calculated a mean interview time of 36 minutes. A unique feature of the SCID-II is that the interview begins with a brief overview which gathers information on behavior and relationships, and provides information about the interviewee's capacity for self-reflection. This allows the interviewer not only to establish rapport, but also allows interviewees to provide a description and consequences of their behavior in their own words (Spitzer et al., 1990). Following the overview, the interview progresses through each relevant disorder.

The format and structure of the SCID-II is very similar to that of the SCID for Axis I disorders. Each page of the interview contains questions, reprinted *DSM-IV* criteria, guidelines for scoring, and space for scores (Spitzer et al., 1990). Initial questions are open-ended and followed up with prompts for elaboration and examples. If further clarification is needed, the interviewer is encouraged to ask supplementary (their own) questions, give examples, present hypothetical situations, and challenge inconsistencies (Spitzer et al., 1990). There are usually two to three interview questions for each personality disorder criterion. In essence, the interviewer is testing diagnostic hypotheses. The ratings are based not on the question response, but an fulfillment of *DSM-IV* criteria. The interviewer is encouraged to use alternate sources of information to assist in rating the criteria, such as observed behavior, medical records, and informants. With slight

modifications, the SCID-II can be administered to an informant (First et al., 1995). Each criterion is rated as one of the following: ? = inadequate information, 1 = symptom clearly absent or criteria not met, 2 = subthreshold condition that almost meets criteria, and 3 = threshold for criteria met.

A rating of "3" is scored only when the interviewee provides convincing, elaborative, and/or exemplary information. A rating of "3" is reserved only for criteria that fulfill the following three guidelines: pathological (outside normal variation), pervasive (occurs in a variety of places and/or with a variety of people), and persistent (occurs with sufficient frequency over at least a five-year period). Specific guidelines for a "3" rating are provided for each criterion in the body of the interview.

Due to the similarity in interview procedures, SCID-II training procedures are almost identical to SCID-I training. As with the SCID-I, clinical judgment is required in the administration and scoring of the SCID-II and thus requires interviewers to have a full understanding of *DSM* nosology and experience in diagnostic interviewing. A user's manual is available for the SCID-II and demonstration videotapes are available. Training workshops can also be arranged with the developers.

4.05.3.3.1 Reliability

The test-retest reliability of the SCID-II was examined within an investigation of the reliability of the Axis I SCID (Williams et al., 1992). In this study, First, Spitzer, Gibbons, Williams, Davies, et al. (1995) administered the SCID-II to 103 psychiatric patients and 181 nonpatients. Two raters independently administered the SCID-II between 1 and 14 days apart. Each SCID-II was preceded by an Axis I SCID evaluation. The SCID-II Personality Questionnaire was given only on the occasion of the first assessment (both SCID-II interviews used the same questionnaire results). Overall weighted kappas were 0.53 and 0.38 for patients and nonpatients, respectively. For the patient sample, kappas were above 0.5 for avoidant, antisocial, paranoid, histrionic, and passive-aggressive personality disorders, and below 0.5 for dependent, self-defeating, narcissistic, borderline, and obsessive-compulsive personality disorders. For the nonpatient sample, kappas were above 0.5 for dependent, histrionic, and borderline personality disorders, and below 0.5 for avoidant, obsessive-compulsive, passive-aggressive, self-defeating, paranoid, and narcissistic personality disorders.

Using the Dutch version of the SCID-II, Arntz et al. (1992) randomly selected 70 mental

health center outpatients and conducted SCID-II interviews. Inter-rater reliability was determined by comparing criteria scores between a primary interviewer and an observer. With the exception of a few criteria, inter-rater reliability for each criterion was good to excellent. Eighty-four of 116 *DSM-III-R* criteria had ICCs higher than 0.75, and 14 had reliability ranging from 0.60 to 0.75. Inter-rater reliability was not able to be rated for 12 criteria due to lack of variance. Inter-rater agreement for specific personality disorders was good with kappas ranging from 0.65 to 1.

Several other studies report good to excellent inter-rater reliability and agreement using the SCID-II (Brooks, Baltazar, McDowell, Munjack, & Bruns, 1991; Fogelson, Nuechterlein, Asarnow, Subotnik, & Talovic, 1991; Malow, West, Williams, & Sutker, 1989; Renneberg, Chambless, & Gracely, 1992). However, these studies contained two or more of the following limitations: restricted number of personality disorders, a homogeneous population and a small number of participants.

4.05.3.3.2 Summary

The SCID-II differs from other personality interviews in several respects. Although other interviews have a disorder-based format available, only the SCID-II has this format as its primary (and only) format of administration. First et al. (1995) maintain that the grouping of questions based on disorder may more closely approximate clinical diagnostic practice and that this grouping forces interviewers to consider criteria in the context of the overarching theme of the disorder. One disadvantage is that the lack of a thematically organized format limits an interviewer's choices, and some have raised concerns that disorder-based organization results in redundancy and repetition with similar items across different diagnoses. Also, the organization of items by disorder may create "halo" effects where a positive criterion rating may bias an interviewer's rating of similar items.

Although the SCID-II screening questionnaire is unusual the IPDE now has a screening questionnaire as well (neither the SIDP-IV or the PDI-IV use screening questionnaires). The SCID-II has shown reliability comparable to other interviews and has been used in a number of studies. The shared format between the SCID-II and the SCID for Axis I disorders should facilitate training on the two measures and may ease the typical requirement that Axis I disorders are assessed and taken into consideration when conducting personality disorder examinations.

4.05.3.4 Personality Disorder Interview-IV

The Personality Disorder Interview-IV (PDI-IV; Widiger, Mangine, Corbitt, Ellis, & Thomas, 1995) is a semistructured interview developed to assess 10 *DSM-IV* personality disorders as well as the two *DSM-IV* personality criteria sets provided for further study (depressive personality disorder, passive-aggressive personality disorder). The PDI-IV is the fourth edition of the Personality Interview Questionnaire (PIQ). The name change was based, in part, on the intent to provide a more descriptive title as the PDI focuses on the assessment of disordered personality.

The PDI-IV provides questions for the assessment of the 94 diagnostic criteria that relate to the 12 *DSM-IV* personality disorders. Criterion ratings are made on a three-point scale (0 = not present, 1 = present according to *DSM-IV* criteria, 2 = present to more severe or substantial degree). Questions from the PDI-IV were selected as useful in determining criterion ratings and additional questions are provided for further elaboration if time allows. However, given the questionnaire's semistructured format, the interviewer may deviate from questions to obtain further information or to address inconsistencies. It is suggested that all positive responses be followed by a request for examples.

The PDI-IV can be administered in a manner either organized by thematic content (as with the IPDE and SIDP) or by *DSM-IV* diagnostic category (as with the SCID-II). Separate interview booklets are provided for these two forms of administration. For occasions when all personality disorders will be assessed, it is recommended that the thematic administration be used. Content areas in the thematic format include Attitudes Towards Self, Attitudes Toward Others, Severity or Comfort with Others, Friendships and Relationships, Conflicts and Disagreements, Work and Leisure, Social Norms, Mood, and Appearance and Perception. The diagnostic format may be preferable when only particular disorders must be assessed. Ratings can be used either to derive categorical or dimensional ratings for *DSM-IV* personality disorders.

The PDI-IV comes with an extensive manual that discusses general issues regarding administration but also provides a thorough discussion of each personality disorder in separate chapters. Each chapter provides an overview of the history and rationale for the personality disorder including discussion of the development of the criterion in *DSM* as well as *ICD* and other criterion sets. Each criterion item is discussed with regard to revisions and rationale for each

item, ratings and questions, and issues and problems relevant to assessing that criterion.

The PDI-IV does not include the use of a self-report questionnaire. However, Widiger et al. (1995) do recommend that a stand-alone self-report inventory be used to assess personality. Scores from such a questionnaire may then be used to inform selection of the most relevant personality disorders to assess on the PDI-IV. Widiger et al. (1995) suggest that the use of such self-report measures will serve the same purpose as screening questionnaires but also will provide more information than measures simply designed for screening purposes.

The PDI-IV manual indicates that, although lay interviewers can administer the PDI-IV, extensive training and supervision is required. Even then, it is recommended that final item scoring be done by an experienced clinician. Ideally, the PDI-IV should be administered and scored by a professional clinician with training and experience in diagnosis and assessment. The PDI-IV manual outlines suggested training beginning with study of the *DSM-IV*, articles on the diagnosis or assessment of personality disorders, and the PDI-IV manual and interview booklets. Following discussion of this literature it is recommended that trainees conduct pilot interviews with nonclinical subjects. Tapes of these initial interviews should be reviewed and feedback provided. It is then suggested that 5–10 patient interviews be conducted and taped for review. Continued taping and systematic review of interviews is recommended to avoid interviewer drift.

4.05.3.4.1 Reliability

Inter-rater agreement for presence vs. absence of personality disorders ranges from 0.37 (histrionic) to 0.81 (antisocial), with a median kappa of 0.65. Agreement for the number of personality disorder criteria met ranges from 0.74 (histrionic, narcissistic, and schizotypal) to 0.90 (obsessive-compulsive and passive-aggressive) and 0.91 (sadistic). Median reliability for the number of PD criteria met was 0.84. Although these data are generally encouraging there are some concerns. Unfortunately, the population on which these reliability data were obtained is not specified nor are the methods for determining rater agreement described. More detailed information may be available from the unpublished dissertation from which these data are derived (Corbitt, 1994).

4.05.3.4.2 Summary

The PDI-IV is built upon the extensive history and experience derived from prior editions of

this interview. The PDI-IV manual is one of the more extensive, thorough, and informative manuals available for the assessment of personality disorders. The flexibility afforded by the choice of either thematic content format or diagnostic category format is also attractive. Despite the accumulation of research on prior versions of the PDI-IV, there is limited reliability data for the PDI-IV. However, the PDI-IV is the only personality interview that has reliability data available for the *DSM-IV* personality disorders.

4.05.4 CHILD AND ADOLESCENT DISORDERS

4.05.4.1 Schedule for Affective Disorders and Schizophrenia for School Age Children

The Schedule for Affective Disorders and Schizophrenia for School Age Children (K-SADS; Puig-Antich & Chambers, 1978) is a semistructured interview designed for research or clinical assessment by a trained clinician. The K-SADS was developed as a child and adolescent version of SADS resulting from research in childhood depression. The K-SADS covers a wide range of childhood disorders but has a strong emphasis on major affective disorders (Roberts, Vargo, & Ferguson, 1989). The interview is intended to assess both past and current episodes of psychopathology in children aged 6–17 years old.

The K-SADS-III-R is compatible with *DSM-III-R* criteria. This version of the SADS provides 31 diagnoses within affective disorders (including depression, bipolar disorder, dysthymia, and cyclothymia), eating disorders, anxiety disorders, behavioral disorders (e.g., conduct disorder, substance abuse/dependence), psychoses, and personality disorders (i.e., schizotypal and paranoid)

The K-SADS is composed of three parts. It begins with an unstructured interview that aims to put the patient at ease and gather information regarding present problems, onset and duration of problems, and treatment history. Following this general interview, the interviewer asks questions relevant to specific symptoms and diagnostic criteria. Sample questions are provided only as a guideline, and modification is encouraged. If initial probe questions are negative, follow-up questions are skipped over. At the conclusion of the interview, observational items are rated (Roberts et al., 1989).

The parent interview should be conducted first, followed by the child interview. The child and parent interview require approximately 90 minutes each. The K-SADS focuses on the last week and most intense symptoms over the last

12 months. Each time period is rated independently and a summary score is made. Diagnostic criteria are rated as present or absent, and then rated on severity (Ambrosini, Metz, Prabucki, & Lee, J., 1989). Ultimately, diagnoses are given based on clinical judgment (Hodges, McKnew, Burbach, & Roebuck, 1987).

As with the SADS, the K-SADS requires extensive training and experience in psychiatric interviewing but has an added burden of conducting interviews with adults (parent/guardian) and children. Full familiarity with *DSM-III-R* is required. Training typically requires viewing videotapes and the conduct of practice interviews with ongoing supervision.

4.05.4.1.1 Reliability

Chambers et al. (1985) examined test–retest reliability of the K-SADS administered to children and parents. Test–retest reliability of major diagnoses was generally adequate with kappas ranging from 0.54 to 0.74, with the exception of anxiety disorder ($K = 0.24$). Individual symptoms and symptom scales generally showed adequate test–retest reliability with anxiety-related symptoms showing the lowest reliability. Agreement between the parent and child interviews varied greatly, ranging from poor to excellent. This later finding suggests the nonredundant aspect of these two interviews.

Inter-rater reliability was examined in videotaped K-SADS-III-R interviews by Ambrosini, et al. (1989). Inter-rater agreement among child, parent, combined interview ratings, and across time frames (present episode and last week) ranged from acceptable ($K = 0.53$) to excellent ($K = 1$) for major depression, minor depression, overanxious disorder, simple phobia, separation anxiety, oppositional, and attention deficit. Of the 36 kappa values, 30 were 0.75 or higher.

Apter, Orvaschel, Laseg, Moses, and Tyano (1989) examined inter-rater and test–retest agreement in a sample of adolescent inpatients (aged 11 to 18 years). Overall inter-rater and test–retest agreement was high with kappas of 0.78. Reliability of symptom scales was also adequate with ICCs of 0.63–0.97 for inter-rater agreement and ICCs of 0.55–0.96 for test–retest agreement. Diagnostic agreement between parent and child interviews (conducted by different clinicians for each informant) was generally low with an overall kappa of 0.42. Parent–child agreement for symptom scales was particularly low for anxiety symptoms.

4.05.4.1.2 Summary

The K-SADS extensively covers the major affective disorders and has adequate coverage of

other childhood disorders. It has been one of the main diagnostic interviews available for use with children and adolescents. Reliability data are very positive for a number of disorders. However, reliability data are largely for *DSM-III* diagnoses and limited data are available for *DSM-III-R* diagnoses (no data are available for *DSM-IV*).

4.05.4.2 Child Assessment Schedule

The Child Assessment Schedule (CAS; Hodges, Kline, Stern, Cytryn, & McKnew, 1982) is a structured interview that is unique in that it is modeled after traditional clinical interviews with children. The interview is organized around thematic topics (e.g., school, friends) with diagnostic items inserted within these topics. Structured questions are provided in a format that is intended to develop rapport. Hodges (1993) has noted that about half of the CAS material related to clinical content does not reflect directly on diagnostic criteria.

The CAS is organized into three parts. In the first part 75 questions are organized into 11 content areas: school, friends, activities and hobbies, family, fears, worries, self-image, mood, somatic concerns, expression of anger, and thought disorder symptoms. Items are rated true (presence of symptom), false (absence of symptom), ambiguous, no response, or not applicable. In the second part the onset and duration of symptoms is assessed. In the third part of the CAS, following completion of the interview, 56 items are rated based on observations during the interview. These items include the following areas: insight, grooming, motor coordination, activity level, other spontaneous physical behaviors, estimate of cognitive ability, quality of verbal communications, quality of emotional expression, and impressions about the quality of interpersonal interactions. A parallel form of the CAS is available for administration to parents (P-CAS). The same inquiries are made, with parents being asked about the child.

Quantitative scales can be obtained for a total score, scores for content areas and for symptom complexes. The internal consistency of the scale scores has been examined and are generally adequate with a few exceptions. Symptom scales have been found to be internally consistent (Hodges, Saunders, Kashani, Hamlett, & Thompson, 1990), especially in a psychiatric sample with some attenuation in medically ill and community samples (particularly for anxiety symptoms). Hodges and Saunders (1989) examined the internal consistency of content scales for both the CAS and P-CAS. For the

CAS, content scales generally had alphas greater than 0.70 but low internal consistency was found for Activities and Reality Functioning. For the P-CAS, content scales with alphas below 0.60 were Activities, Reality-testing Symptoms, Self-image, and Fears. Diagnoses for *DSM-III-R* can be derived in addition to these scale scores. The CAS takes approximately 45 minutes to one hour to complete.

It is recommended that the CAS be administered by trained clinicians (although lay interviewers have been used; Hodges et al., 1987). Guidelines for administering, scoring, and interpreting the CAS are contained in the CAS manual (Hodges, 1985) and in guidelines established for achieving rater reliability (Hodges, 1983).

4.05.4.2.1 Reliability

In the initial rater reliability study, Hodges, McKnew, Cytryn, Stern and Kline (1982) examined inter-rater agreement using videotaped interviews. For symptom scores, mean kappas were generally close to or exceeded 0.60. Mean correlations across raters for content areas was 0.63 or above with the exception of worries (0.59). For symptom complexes mean correlations were 0.68 or above except for attention deficit without hyperactivity (0.58), separation anxiety (0.56), and socialized conduct (0.44). Hodges et al. (1982) also report inter-rater reliability on a small sample ($N = 10$). Correlations for items, content scores, and symptom complex scores were all above 0.85. Verhust, Althaus, and Berden (1987) also have reported inter-rater reliability for contents scores using a small number ($N = 10$) of videotaped interviews. Correlations for content areas ranged from 0.70 to 0.97.

In the only test-retest reliability study of the CAS, Hodges, Cools, and McKnew (1989) examined agreement over a mean of five days with an inpatient sample. Intraclass correlations indicated good reliability for the total CAS score and scale scores. Kappas for *DSM-III* diagnoses of conduct disorder, depression, and anxiety were above 0.70. However, the kappa for attention deficit disorder was only 0.43.

The concordance between the CAS and the K-SADS was examined by Hodges et al. (1987). Lay interviewers were used and agreement was examined for both child and parent interviews for four major diagnostic categories (attention deficit disorder, conduct disorders, anxiety disorders, and affective disorders). Only present episodes were evaluated. Child only diagnostic concordance between the CAS and K-SADS was poor for attention deficit disorder and anxiety disorders (kappas less than 0.40). Better

agreement was obtained for parent only interviews or in combinations of child and parent interviews. Anxiety disorders generally had low concordance across informant conditions.

The concordance between child and parent interviews has also been examined with the CAS. Verhulst et al. (1987) found low to moderate correlations between parent- and child-derived content areas, somatic concerns, and observational judgments. Of 22 correlations, only four exceeded 0.40. The total score correlation was 0.58, indicating approximately 34% shared variance. Hodges, Gordon, and Lennon (1990) also found low to moderate correlations between parent and child interview ratings. For diagnostic areas, the lowest correlations (those below 0.30) were obtained for overanxious disorder, oppositional disorder, and separation anxiety. Low correlations (again below 0.30) were also found in the content areas of fears, worries and anxieties, and physical complaints. These data indicate reasonable parent-child agreement for conduct/behavioral problems, moderate agreement for affective symptoms, and low agreement for anxiety symptoms. As with other child assessment measures, the greatest parent-child agreement appears to be for observable behavior and the lowest for subjective experiences (Hodges et al., 1990; Hodges, 1993).

4.05.4.2.2 Summary

The CAS appears to provide a reliable assessment of a range of symptoms and shows reasonable convergence with noninterview measures. It does not cover a number of disorders including sleep disorders, eating disorders, alcohol or drug use disorders, or mania. Although it provides a broad clinical assessment, some users may find that the presence of many CAS items that do not reflect directly on diagnostic criteria is inefficient. Inter-rater agreement for diagnoses studied appears adequate. However, only one small-scale study has examined test-retest reliability for a subset of diagnoses. No reliability data are available for *DSM-IV* diagnoses.

4.05.4.3 Child and Adolescent Psychiatric Assessment

The Child and Adolescent Psychiatric Assessment (CAPA; Angold et al., 1995) was developed in order to assess a wide range of diagnostic classifications including *DSM-III*, *DSM-III-R*, *ICD-9*, and *ICD-10*. Additionally, other symptoms of clinical interest are evaluated. As with other interviews, the CAPA can be

administered to children and parents. The CAPA has four sections, three of which pertain to the interview proper. The time period addressed is the three months preceding the interview. In the Introduction, the interview is conducted in a conversational manner in order to establish rapport. Questions within the Introduction address three areas: home and family life, school life, peer groups and spare time activities.

The second section is the Symptom Review which has a disorder-based organization. A wide range of disorders are covered including anxiety disorders, obsessive-compulsive disorders, depressive disorders, manic disorders, somatization disorders, food-related disorders, sleep disorders, elimination disorders, tic disorders and trichotillomania, disruptive behavior disorders, tobacco use, alcohol, psychotic disorders, life events and post-traumatic stress disorder (PTSD), and drugs. Due to problems in child report with some disorders, only the parent interview assesses sleep terror disorder, sleepwalking disorder, and attention deficit hyperactivity disorder. Alternatively, because parents may be a poor source of information for children's substance use, delusions, hallucinations, and thought disorder these items are abbreviated in the parent interview with more extensive coverage in the child interview.

The third section of the interview assesses incapacity. At this point the interviewer reviews symptom information and questions about the effects of symptoms in 17 areas of psychosocial impairment. Impairment is evaluated in the three domains of home and family life, school life, peer groups and spare time activities. Finally, following the interview, observations of interview behavior are rated for 67 items. These items cover level of activity, child's mood state, quality of child's social interaction during interview, and psychotic behavior.

Detailed questions are provided for each interview item in the CAPA. There are three levels of questions. Screening questions allow a skip-out of a section. If the screen question is positive two levels of probes are provided. Emphasized probes are required and should be asked for all subjects. Discretionary probes are provided if further information is required. A glossary is provided to be used in conjunction with the standardized questions. The glossary provides operational definitions of symptom items. These definitions were based on a review of several of the existing clinical child interviews. The glossary also provides explicit rating principles including a formal definition of the item, ratings of intensity (from 0, absent, to 3, present at a higher intensity level), duration, length of time symptom is occurring,

and psychosocial impairment related to the symptom.

A wealth of information is obtained with the CAPA. Fortunately, a computer program is available in order to summarize these data with a series of diagnostic algorithms (the CAPA Originated Diagnostic Algorithms; CODA). The CODA can generate diagnoses according to *DSM-III*, *DSM-III-R*, *DSM-IV*, and *ICD-10* systems as well as symptom scores for particular diagnostic areas.

Angold et al. (1995) report that the CAPA has been used with a variety of populations (both clinical and general population) in both the UK and the USA. Training requires four to five weeks with emphasis on practice administering the CAPA and group ratings of tapes. Based on its use in multiple clinical centers explicit training criteria have been developed and details about the CAPA and its training requirements can be obtained from Angold.

4.05.4.3.1 Reliability

Angold and Costello (1995) examined test-retest reliability in a clinical sample. Interviews were conducted with children only and were completed within 11 days of each other. Kappas for specific *DSM-III-R* diagnoses were all above 0.73, with the exception of conduct disorder ($K = 0.55$). Agreement on symptom scales for these disorders was also high with ICCs above 0.60 except for conduct disorder ($ICC = 0.50$). No reliability data were available for a number of disorders covered in the CAPA including obsessive-compulsive disorders, manic disorders, food-related disorders, sleep disorders, elimination disorders, tic disorders, psychotic disorders, or life events and PTSD.

4.05.4.3.2 Summary

The CAPA appears to offer a thorough clinical evaluation that incorporates several diagnostic criteria. It provides a broader assessment with more contemporary diagnostic nosology than some other instruments. However, this breadth of assessment does come at a cost. The CAPA administered to the child alone can take one hour and coding can take another 45 minutes. The CAPA is not recommended for use with children under the age of eight. Additionally, the CAPA is limited to the three months preceding the interview. Although reliability data are encouraging, these data are limited to child only interviews, are not available for a number of disorders covered by the CAPA, and are not available for *DSM-IV* diagnoses. It will be important to determine

the reliability of other disorders as well as that of parent interviews, and whether diagnostic agreement is improved with both child and parent administration.

4.05.4.4 Diagnostic Interview Schedule for Children

The Diagnostic Interview Schedule for Children (DISC) is a highly structured interview designed to assess most child and adolescent psychiatric disorders (Jensen et al., 1995). The interview was introduced in 1982 as a child version of the Diagnostic Interview Schedule (DIS). The DISC was intended to be administered by lay interviewers and used for epidemiological research (Shaffer et al., 1993). The version current in the late 1990s, DISC-2.1 covers 35 diagnostic criteria for the *DSM-III-R*, and contains a child (DISC-C) and parent (DISC-P) interview. The DISC was designed for children and adolescents ranging from 6 to 18 years old.

DISC interviewers are encouraged not to deviate from the order, wording, and scoring procedures. The child and parent interviews of the DISC-2.1 require approximately 60–75 minutes each (Jensen et al., 1995). Questions, organized into six separate diagnostic modules, inquire about current and past symptoms, behaviors, and emotions of most child and adolescent diagnoses. Diagnostic criteria are initially assessed with a broad “stem question” (with a low diagnostic threshold) and, if endorsed, followed with “contingent questions” to determine criteria requirements, duration, frequency, impairment, and other modifiers (Fisher et al., 1993). The DISC-2.1 focuses on the last six months and a graphic timeline is used to assist in recall (Fisher et al., 1993; Jensen et al., 1995). At the end of each module, supplementary questions are provided to assess onset, current impairment, treatment history, and precipitating stressors. Questions are rated as: “no,” “yes,” or “sometimes” or “somewhat,” and a computer algorithm generates diagnoses.

The DISC was specifically developed for use by lay interviewers in epidemiological research. Interviewer training typically takes one to two weeks. No differences in performance have been found between clinicians and lay interviewers using the DISC-1 (Shaffer et al., 1993). A user’s manual for the DISC is available.

4.05.4.4.1 Reliability

Jensen et al. (1995) examined test-retest reliability in both a clinical and community sample across three sites. In the clinic sample,

for major diagnostic categories, test–retest agreement was adequate for parents (K range = 0.58–0.70) and was generally higher than that obtained for child interviews (K range = 0.39–0.86). Using a combined diagnostic algorithm, test–retest agreement was adequate (K range = 0.50–0.71). Inter-rater agreement was lower for the community sample with test–retest agreement lower for parents (K range = 0.66) and children (K range = 0.23–0.60). The combined diagnostic algorithm for the community sample continued to provide low agreement (K range = 0.26–0.64). Instances of diagnostic disagreement in the clinic sample appeared to be related to an absolute decrease in the number of symptoms at the time of the second interview. Low reliability in the community sample was attributed to decreased symptom severity, the presence of threshold cases, and other unknown factors.

Other studies have generally found adequate test–retest reliability for the DISC. One general pattern that has emerged is greater agreement when examining parent interviews. Schwab-Stone et al. (1993) interviewed 41 adolescents (aged 11–17 years) and 39 parents twice (1–3 weeks apart) with the DISC-R. Inter-rater agreement ranged from poor (K = 0.16) to good (K = 0.77) for the child interviews, and ranged from fair (K = 0.55) to excellent (K = 0.88) for the parent interviews. Schwab-Stone, Fallon, Briggs, and Crowther (1994) interviewed 109 preadolescents (aged 6–11 years) and their parents twice (7–18 days apart) with the DISC-R. Inter-rater agreement ranged from poor (K = 0) to fair (K = 0.56) for the child interviews, and from poor (K = 0.43) to excellent (K = 0.81) for the parent interviews. Based on the lower inter-rater agreement for preadolescents, Schwab-Stone et al. (1994) concluded that highly structured interviews were not appropriate for directly assessing young children due to lower endorsement of symptoms and unreliable reporting within the interview.

Most of the reliability studies on the DISC have examined only the most common childhood diagnoses and little information is available on uncommon diagnoses. From a clinical sample of relatively uncommon diagnoses, Fisher et al. (1993) interviewed 93 children (aged 8–19 years) and 75 parents with the DISC-2.1. Using the clinic diagnosis as a standard, the DISC-2.1 had good (0.73) to excellent (1) sensitivity in identifying eating disorders, major depressive disorders, obsessive-compulsive disorder, psychosis, substance use disorders, and tic disorders. The DISC-2.1 was noted to be less sensitive for major depressive disorder than other interviews (K-SADS, DICA, CAS).

4.05.4.4.2 Summary

The DISC's design for epidemiological research constitutes several advantages. First, the highly structured interview may be administered by nonclinicians. Second, the DISC contains the full range of disorders. Finally, the DISC has been thought to contain a lower threshold for disorders than other interviews, which makes it ideal for screening and use in the general population (Roberts et al., 1989). However, the design of the DISC has several disadvantages. It may be too restrictive, at times not allowing the interviewer to probe further and adapt the interview to accommodate special situations. The DISC has shown to be unreliable among young children and is fairly long requiring 60–75 minutes each for the two interviews. Research on the concordance between the DISC and clinical structured interviews such as the K-SADS has not been examined.

4.05.4.5 Diagnostic Interview for Children and Adolescents

The Diagnostic Interview for Children and Adolescents (DICA) is a highly structured interview designed to be used by lay interviewers for clinical and epidemiological research. The interview assesses the present episode of a wide range of psychopathology among children aged 6–17 years (Roberts et al., 1989). The interview initially appeared in 1969, and was revised in 1981 to emulate the organization of the DIS and based on *DSM-III* criteria (Welner, Reich, Herjanic, Jung, & Amado, 1987). The DICA was subsequently revised to conform to *DSM-III-R* diagnoses (DICA-R; Kaplan & Reich, 1991). In addition to coverage of *DSM-III-R*, the DICA-R was also revised so that questions were presented in a more conversational style.

The DICA-R is organized into 15 sections. Two sections cover general and sociodemographic information, 11 sections relate to disorders and symptomatology, and remaining sections address menstruation, psychosocial stressors, and clinical observations. The DICA consists of a separate parent (DICA-P) and child (DICA-C) interview. The child interview requires 30–40 minutes to administer, while the parent interview takes longer due to additional questions on developmental history, medical history, socioeconomic status, and family history (Roberts et al., 1989). For each diagnostic criterion, one or more questions elicit information. Follow-up questions are skipped if primary questions are responded negatively. Responses on the DICA-R are coded on a four-point scale: “no,” “rarely,”

“sometimes” or “somewhat,” and “yes.” Following each diagnostic section, specific *DSM* criteria are listed to assist in deriving diagnoses (Welner et al., 1987).

4.05.4.5.1 Reliability

Limited data on the inter-rater agreement of the DICA is available. Only one study has provided data pertaining to individual diagnoses with an adequate description of sample and methods. Welner et al. (1987) administered two independent interviews (1–7 days apart) to 27 psychiatric inpatients (aged 7–17 years). Using lay interviewers, inter-rater agreement was excellent across diagnostic categories (*K* range = 0.76–1, median = 0.86).

Similar to other interviews, diagnoses derived from the parent and child interviews vary. Welner et al. (1987) examined concordance between child and parent interviews among 84 outpatients (ages 7–17 years). Fair to excellent concordance was noted (*K* range = 0.49–0.80, median = 0.63). However, other studies have found more modest concordance between parent and child interviews with median kappas below 0.30 (Earls, Reich, Jung, & Cloninger, 1988; Sylvester, Hyde, & Reichler, 1987).

4.05.4.5.2 Summary

The DICA-R appears to be a well-developed instrument that has taken special care in the writing and sequencing of questions. Although the DICA has been used in a number of studies, only a limited amount of reliability information is available. No reliability information is available for the DICA-R. Other child and adolescent interviews may be more attractive because of the relatively greater inter-rater reliability information.

4.05.5 SUMMARY

There has been an enormous amount of research conducted on the development and use of structured clinical interviews since the late 1960s. This research has yielded diagnostic interviews that address an array of clinical diagnoses in both adult and child populations. The use of structured interviews can not only provide reliable diagnostic evaluations but can also serve to ensure a broad and thorough clinical assessment. Although most readily embraced in research settings, it is anticipated (and hoped) that structured diagnostic interviews will become more commonplace in clinical applied settings as part of the standard assessment tools that clinicians use regularly.

4.05.6 REFERENCES

- Ambrosini, P. J., Metz, C., Prabucki, K., & Lee, J. (1989). Videotape reliability of the third revised edition of the K-SADS. *Journal of the American Academy of Child and Adolescent Psychiatry*, 28, 723–728.
- American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- Andreasen, N. C. (1983). *The Scale for the Assessment of Negative Symptoms (SANS)*. Iowa City, IA: The University of Iowa.
- Andreasen, N. C. (1984). *The scale for the assessment of positive symptoms (SAPS)*. Iowa City, IA: The University of Iowa.
- Andreasen, N. C. (1987). *Comprehensive assessment of symptoms and history*. Iowa City, IA: The University of Iowa.
- Andreasen, N. C., Flaum, M., & Arndt, S. (1992). The comprehensive assessment of symptoms and history (CASH): An instrument for assessing diagnosis and psychopathology. *Archives of General Psychiatry*, 49, 615–623.
- Andreasen, N. C., Grove, W. M., Shapiro, R. W., Keller, M. B., Hirschfeld, R. M. A., & McDonald-Scott, P. (1981). Reliability of lifetime diagnoses: A multicenter collaborative perspective. *Archives of General Psychiatry*, 38, 400–405.
- Angold, A., & Costello, E. J., (1995). A test–retest study of child-reported psychiatric symptoms and diagnoses using the Child and Adolescent Psychiatric Assessment (CAPA-C). *Psychological Medicine*, 25, 755–762.
- Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., & Rutter, M. (1995). The Child and Adolescent Psychiatric Assessment (CAPA). *Psychological Medicine*, 25, 739–753.
- Anthony, J. C., Folstein, M., Romanoski, A. J., Von Korff, M. R., Nestadt, G. R., Chahal, R., Merchant, A., Brown, H., Shapiro, S., Kramer, M., & Gruenberg, E. M. (1985). Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis: Experience in eastern Baltimore. *Archives of General Psychiatry*, 42, 667–675.
- Apter, A., Orvaschel, H., Laseg, M., Moses, T., & Tyano, S. (1989). Psychometric properties of the K-SADS-P in an Israeli adolescent inpatient population. *Journal of the American Academy of Child and Adolescent Psychiatry*, 28, 61–65.
- Arntz, A., van Beijsterveldt, B., Hoekstra, R., Hofman, A., Eussen, M., & Sallaerts, S. (1992). The inter-rater reliability of a Dutch version of the Structured Clinical Interview for DSM-III-R personality disorders. *Acta Psychiatrica Scandinavica*, 85, 394–400.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1962). Reliability of psychiatric diagnoses: 2. A study of consistency of clinical judgments and ratings. *American Journal of Psychiatry*, 119, 351–357.
- Bromet, E. J., Dunn, L. O., Connell, M. M., Dew, M. A., & Schulberg, H. C. (1986). Long-term reliability of diagnosing lifetime major depression in a community sample. *Archives of General Psychiatry*, 43, 435–440.
- Brooks, R. B., Baltazar, P. L., McDowell, D. E., Munjack, D. J., & Bruns, J. R. (1991). Personality disorders co-occurring with panic disorder with agoraphobia. *Journal of Personality Disorders*, 5, 328–336.
- Chambers, W. J., Puig-Antich, J., Hirsch, M., Paez, P., Ambrosini, P. J., Tabrizi, M. A., & Davies, M. (1985). The assessment of affective disorders in children and adolescents by semistructured interview. *Archives of General Psychiatry*, 42, 696–702.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

- Cooper, J. E., Copeland, J. R. M., Brown, G. W., Harris, T., & Gurlay, A. J. (1977). Further studies on interviewer training and inter-rater reliability of the Present State Exam (PSE). *Psychological Medicine*, 7, 517-523.
- Cooper, J. E., Kendell, R. E., Gurland, B. J., Sharpe, L., Copeland, J. R. M., & Simon, R. (1972). *Psychiatric diagnosis in New York and London. Maudsley monographs*. London: Oxford University Press.
- Corbitt, E. M. (1994). *Sex bias and the personality disorders: A reinterpretation from the five-factor model*. Unpublished doctoral dissertation, University of Kentucky, Lexington.
- Cottler, L. B. (1991). The CIDI and CIDI-Substance Abuse Module (SAM): Cross-cultural instruments for assessing DSM-III, DSM-III-R and ICD-10 criteria. *Research Monographs*, 105, 220-226.
- Earls, R., Reich, W., Jung, K. G., & Cloninger, C. R. (1988). Psychopathology in children of alcoholic and antisocial parents. *Alcoholism: Clinical and Experimental Research*, 12, 481-487.
- Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry*, 35, 837-844.
- Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, 33, 766-771.
- Erdman, H. P., Klein, M. H., Greist, J. H., Bass, S. M., Bires, J. K., & Machtinger, P. E. (1987). A comparison of the Diagnostic Interview Schedule and clinical diagnosis. *American Journal of Psychiatry*, 144, 1477-1480.
- Escobar, J. I., Randolph, E. T., Asamen, J., & Karno, M. (1986). The NIMH-DIS in the assessment of DSM-III schizophrenic disorder. *Schizophrenia Bulletin*, 12, 187-194.
- Faraone, S. V., Blehar, M., Pepple, J., Moldin, S. O., Norton, J., Nurnberger, J. I., Malaspina, D., Kaufman, C. A., Reich, T., Cloninger, C. R., DePaulo, J. R., Berg, K., Gershon, E. S., Kirch, D. G., & Tsuang, M. T. (1996). Diagnostic accuracy and confusability analyses: An application to the Diagnostic Interview for Genetic Studies. *Psychological Medicine*, 26, 401-410.
- Farmer, A. E., Katz, R., McGuffin, P., & Bebbington, P. (1987). A comparison between the Present State Examination and the Composite International Interview. *Archives of General Psychiatry*, 44, 1064-1068.
- Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry*, 26, 57-63.
- First, M. B., Gibbon, M., Spitzer, R. L., & Williams, J. B. W. (1996). *User's guide for the Structured Clinical Interview for DSM-IV Axis I Disorders-Research Version (SCID-I, version 2.0, February 1996 Final version)*. New York: Biometrics Research Department, New York State Psychiatric Institute.
- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). Part I: Description. *Journal of Personality Disorders*, 9, 83-91.
- First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., & Rounsaville, B. (1995). The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). Part II: Multi-site test-retest reliability study. *Journal of Personality Disorders*, 9, 92-104.
- Fisher, P. W., Shaffer, D., Piacentini, J. C., Lapkin, J., Kafantaris, V., Leonard, H., & Herzog, D. B. (1993). Sensitivity of the Diagnostic Interview Schedule for Children, 2nd Edition (DISC-2.1) for specific diagnoses of children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 32, 666-673.
- Fogelson, D. L., Nuechterlein, K. H., Asarnow, R. F., Subotnik, K. L., & Talovic, S. A. (1991). Inter-rater reliability of the Structured Clinical Interview for DSM-III-R, Axis II: schizophrenia spectrum and affective spectrum disorders. *Psychiatry Research*, 39, 55-63.
- Folstein, M. F., Folstein, S. E., & McHugh, P. (1975). "Mini Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189-198.
- Ford, J., Hillard, J. R., Giesler, L. J., Lassen, K. L., & Thomas, H. (1989). Substance abuse/mental illness: Diagnostic issues. *American Journal of Drug and Alcohol Abuse*, 15, 297-307.
- Fyer, A. J., Mannuzza, S., Martin, L. Y., Gallops, M. S., Endicott, J., Schleyer, B., Gorman, J. M., Liebowitz, M. R., & Klein, D. F. (1989). Reliability of anxiety assessment, II: Symptom assessment. *Archives of General Psychiatry*, 46, 1102-1110.
- Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, 38, 408-413.
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23, 56-62.
- Helzer, J. E., Robins, L. N., McEvoy, L. T., Spitznagel, E. L., Stolzman, R. K., Farmer, A., & Brockington, I. F. (1985). A comparison of clinical and diagnostic interview schedule diagnoses: Physician reexamination of lay-interviewed cases in the general population. *Archives of General Psychiatry*, 42, 657-666.
- Hodges, K. (1983). *Guidelines to aid in establishing inter-rater reliability with the Child Assessment Schedule*. Unpublished manuscript.
- Hodges, K. (1985). *Manual for the Child Assessment Schedule*. Unpublished manuscript.
- Hodges, K. (1993). Structured interviews for assessing children. *Journal of Child Psychology and Psychiatry*, 34, 49-68.
- Hodges, K., Cools, J., & McKnew, D. (1989). Test-retest reliability of a clinical research interview for children: The Child Assessment Schedule (CAS). *Psychological Assessment: Journal of Consulting and Clinical Psychology*, 1, 317-322.
- Hodges, K., Gordon, Y., & Lennon, M. P. (1990). Parent-child agreement on symptoms assessed via a clinical research interview for children: The Child Assessment Schedule (CAS). *Journal of Child Psychology and Psychiatry*, 31, 427-436.
- Hodges, K., Kline, J., Stern, L., Cytryn, L., & McKnew, D. (1982). The development of a child assessment interview for research and clinical use. *Journal of Abnormal Child Psychology*, 10, 173-189.
- Hodges, K., McKnew, D., Burbach, D. J., & Roebuck, L. (1987). Diagnostic concordance between the Child Assessment Schedule (CAS) and the Schedule for Affective Disorders and Schizophrenia for School-age Children (K-SADS) in an outpatient sample using lay interviewers. *Journal of the American Academy of Child and Adolescent Psychiatry*, 26, 654-661.
- Hodges, K., McKnew, D., Cytryn, L., Stern, L., & Kline, J. (1982). The Child Assessment Schedule (CAS) diagnostic interview: A report on reliability and validity. *Journal of the American Academy of Child Psychiatry*, 21, 468-473.
- Hodges, K., & Saunders, W. (1989). Internal consistency of a diagnostic interview for children: The Child Assessment Schedule. *Journal of Abnormal Child Psychology*, 17, 691-701.
- Hodges, K., Saunders, W. B., Kashani, J., Hamlett, K., & Thompson, R. J. (1990). *Journal of the American*

- Academy of Child and Adolescent Psychiatry*, 29, 635–641.
- Hyler, S. E., Skodol, A. E., Kellman, H. D., Oldham, J. M., & Rosnik, L. (1990). Validity of the Personality Diagnostic Questionnaire-Revised: Comparison with two structured interviews. *American Journal of Psychiatry*, 147, 1043–1048.
- Hyler, S. E., Skodol, A. E., Oldham, J. M., Kellman, D. H., & Doldge, N. (1992). Validity of the Personality Diagnostic Questionnaire-Revised: A replication in an outpatient sample. *Comprehensive Psychiatry*, 33, 73–77.
- Jackson, H. J., Gazis, J., Rudd, R. P., & Edwards, J. (1991). Concordance between two personality disorder instruments with psychiatric inpatients. *Comprehensive Psychiatry*, 32, 252–260.
- Janca, A., Robins, L. N., Cottler, L. B., & Early, T. S. (1992). Clinical observation of assessment using the Composite International Diagnostic Interview (CIDI): An analysis of the CIDI field trials—Wave II at the St Louis Site. *British Journal of Psychiatry*, 160, 815–818.
- Jensen, P., Roper, M., Fisher, P., Piacentini, J., Canino, G., Richters, J., Rubio-Stipec, M., Dulcan, M., Goodman, S., Davies, M., Rae, D., Shaffer, D., Bird, H., Lahey, B., & Schwab-Stone, M. (1995). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). *Archives of General Psychiatry*, 52, 61–71.
- Kaplan, L. M., & Reich, W. (1991). *Manual for Diagnostic Interview for Children and Adolescents-Revised (DICA-R)*. St Louis, MO: Washington University.
- Keller, M. B., Lavori, P. W., McDonald-Scott, P., Scheftner, W. A., Andreasen, N. C., Shapiro, R. W., & Croughan, J. (1981). Reliability of lifetime diagnoses and symptoms in patients with a current psychiatric disorder. *Journal of Psychiatric Research*, 16, 229–240.
- Kendell, R. E., Everitt, B., Cooper, J. E., Sartorius, N., & David, M. E. (1968). Reliability of the Present State Examination. *Social Psychiatry*, 3, 123–129.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Loranger, A. W., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., Coid, B., Dahl, A., Diekstra, R. F. W., Ferguson, B., Jacobsberg, L. B., Janca, A., Mombour, W., Pull, C., Ono, Y., Regier, D. A., Sartorius, N., & Sumba R. O. (1995). *The International Personality Disorder Examination (IPDE) manual*. New York: World Health Organization.
- Loranger, A. W., Lenzenweger, M. F., Gartner, A. F., Susman, V. L., Herzig, J., Zammit, G. K., Gartner, J. D., Abrams, R. C., & Young, R. C. (1991). Trait-state artifacts and the diagnosis of personality disorders. *Archives of General Psychiatry*, 48, 720–728.
- Loranger, A. W., Sartorius, N., Andreoli, A., Berger, P., Buchheim, P., Channabasavanna, S. M., Coid, B., Dahl, A., Diekstra, R. F. W., Ferguson, B., Jacobsberg, L. B., Mombour, W., Pull, C., Ono, Y., & Regier, D. A. (1994). The international personality disorder examination. *Archives of General Psychiatry*, 51, 215–224.
- Luria, R. E., & Berry, R. (1979). Reliability and descriptive validity of PSE syndromes. *Archives of General Psychiatry*, 36, 1187–1195.
- Luria, R. E., & Berry, R. (1980). Teaching the Present State Examination in American. *American Journal of Psychiatry*, 137, 26–31.
- Luria, R. E., & McHugh, P. R. (1974). Reliability and clinical utility of the “Wing” Present State Examination. *Archives of General Psychiatry*, 30, 866–971.
- Malow, R. M., West, J. A., Williams, J. L., Sutker P. B. (1989). Personality disorders classification and symptoms in cocaine and opioid addicts. *Journal of Consulting and Clinical Psychology*, 57, 765–767.
- Mannuzza, S., Fyer, A. J., Martin, L. Y., Gallops, M. S., Endicott, J., Gorman, J., Liebowitz, M. R., & Klein, D. F. (1989). Reliability of anxiety assessment, I: Diagnostic agreement. *Archives of General Psychiatry*, 46, 1093–1101.
- Matarazzo, J. D. (1983). The reliability of psychiatric and psychological diagnosis. *Clinical Psychology Review*, 3, 103–145.
- McDonald-Scott, P., & Endicott, J. (1984). Informed versus blind: The reliability of cross-sectional ratings of psychopathology. *Psychiatry Research*, 12, 207–217.
- McGuffin, P., & Farmer, A. E. (1990). *Operational Criteria (OPCRIT) Checklist. Version 3.0*. Cardiff, UK: University of Wales.
- McGuffin, P., Katz, R., & Aldrich, J. (1986). Past and Present State Examination: The assessment of “lifetime ever” psychopathology. *Psychological Medicine*, 16, 461–465.
- Nurnberger, J. I., Blehar, M. C., Kaufman, C. A., York-Cooler, C., Simpson, S. G., Harkavy-Friedman, J., Severe, J. B., Malaspina, D., Reich, T., & collaborators from the NIMH Genetics Initiative (1994). Diagnostic Interview for Genetic Studies: Rationale, unique features, and training. *Archives of General Psychiatry*, 51, 849–859.
- O’Boyle, M., & Self, D. (1990). A comparison of two interviews for DSM-III-R personality disorders. *Psychiatry Research*, 32, 85–92.
- Okasha, A., Sadek, A., Al-Haddad, M. K., & Abdel-Mawgoud, M. (1993). Diagnostic agreement in psychiatry: A comparative study between ICD-9, ICD-10, and DSM-III-R. *British Journal of Psychiatry*, 162, 621–626.
- Overall, J., & Gorham, D. (1962). Brief Psychiatric Rating Scale. *Psychological Reports*, 10, 799–812.
- Page, A. C. (1991). An assessment of structured diagnostic interviews for adult anxiety disorders. *International Review of Psychiatry*, 3, 265–278.
- Pfohl, B., Black, D. W., Noyes, R., Coryell, W. H., & Barrash, J. (1990). Axis I/Axis II comorbidity findings: Implications for validity. In J. Oldham (Ed.), *Axis II: New perspectives on validity* (pp. 147–161). Washington, DC: American Psychiatric Association.
- Pfohl, B., Blum, N., & Zimmerman, M. (1995). *The Structured Interview for DSM-IV Personality Disorders (SIDP-IV)*. Iowa City, IA: University of Iowa College of Medicine.
- Pfohl, B., Blum, N., Zimmerman, M., & Stangl, D. (1989). *Structured Interview for DSM-III-R Personality Disorders (SIDP-R)*. Iowa City, IA: University of Iowa College of Medicine.
- Pilkonis, P. A., Heape, C. L., Proietti, J. M., Clark, S. W., McDavid, J. D., & Pitts, T. E. (1995). The reliability and validity of two structured diagnostic interviews for personality disorders. *Archives of General Psychiatry*, 52, 1025–1033.
- Puig-Antich, J., & Chambers, W. J. (1978). *Schedule for Affective Disorders and Schizophrenia for School-age Children: Kiddie SADS (K-SADS)*. New York: Department of Child and Adolescent Psychiatry, New York State Psychiatric Institute.
- Renneberg, B., Chambless, D. L., & Gracely, E. J. (1992). Prevalence of SCID-diagnosed personality disorders in agoraphobic outpatients. *Journal of Anxiety Disorders*, 6, 111–118.
- Rice, J. P., Rochberg, N., Endicott, J., Lavori, P. W., & Miller, C. (1992). Stability of psychiatric diagnoses: An application to the affective disorders. *Archives of General Psychiatry*, 49, 824–830.
- Roberts, N., Vargo, B., & Ferguson, H. B. (1989). Measurement of anxiety and depression in children and adolescents. *Psychiatric Clinics of North America*, 12, 837–860.
- Robins, L. N., Helzer, J. E., Croughan, J., Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and

- validity. *Archives of General Psychiatry*, 38, 381–389.
- Robins, L. N., Helzer, J. E., Ratcliff, K. S., & Seyfried, W. (1982). Validity of the diagnostic interview schedule, version II: DSM-III diagnoses. *Psychological Medicine*, 12, 855–870.
- Robins, L. N., Wing, J., Wittchen, H.-U., Helzer, J. E., Babor, T. F., Burke, J., Farmer, A., Jablenski, A., Pickens, R., Regier, D. A., Sartorius, N., & Towle, L. H. (1988). The Composite International Diagnostic Interview: An epidemiological instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry*, 45, 1069–1077.
- Rodgers, B., & Mann, S. (1986). The reliability and validity of PSE assessments by lay interviewers: A national population survey. *Psychological Medicine*, 16, 689–700.
- Schwab-Stone, M., Fallon, T., Briggs, M., & Crowther, B. (1994). Reliability of diagnostic reporting for children aged 6–11 years: A test–retest study of the Diagnostic Interview Schedule for Children-Revised. *American Journal of Psychiatry*, 151, 1048–1054.
- Schwab-Stone, M., Fisher, P., Piacentini, J., Shaffer, D., Davies, M., & Briggs, M. (1993). The Diagnostic Interview Schedule for Children-Revised version (DISC-R): II. Test–retest reliability. *Journal of the American Academy of Child and Adolescent Psychiatry*, 32, 651–657.
- Segal, D. L., Hersen, M., & Van Hasselt, V. B. (1994). Reliability of the structured clinical interview for DSM-III-R: An evaluative review. *Comprehensive Psychiatry*, 35, 316–327.
- Shaffer, D., Schwab-Stone, M., Fisher, P., Cohen, P., Piacentini, J., Davies, M., Connors, C. K., & Regier, D. (1993). The Diagnostic Interview Schedule for Children-Revised version (DISC-R): I. Preparation, field testing, inter-rater reliability, and acceptability. *Journal of the American Academy of Child and Adolescent Psychiatry*, 32, 643–650.
- Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, 44, 172–177.
- Spengler, P. A., & Wittchen, H. -U. (1988). Procedural validity of standardized symptom questions for the assessment of psychotic symptoms: A comparison of the DIS with two clinical methods. *Comprehensive Psychiatry*, 29, 309–322.
- Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*, 24, 399–411.
- Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry*, 17, 83–87.
- Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research Diagnostic Criteria: Rationale and reliability. *Archives of General Psychiatry*, 35, 773–782.
- Spitzer, R. L., & Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *British Journal of Psychiatry*, 125, 341–347.
- Spitzer, R. L., Fleiss, J. L., & Endicott, J. (1978). Problems of classification: Reliability and validity. In M. A. Lipton, A. DiMascio, & K. F. Killam (Eds.), *Psychopharmacology: A generation of progress* (pp. 857–869). New York: Raven Press.
- Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1990). *User's guide for the Structured Clinical Interview for DSM-III-R (SCID)*. Washington, DC: American Psychiatric Press.
- Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1992). The Structured Clinical Interview for DSM-III-R (SCID): I: History, rationale, and description. *Archives of General Psychiatry*, 49, 624–629.
- Standage, K. (1989). Structured interviews and the diagnosis of personality disorders. *Canadian Journal of Psychiatry*, 34, 906–912.
- Stangl, D., Pfohl, B., Zimmerman, M., Bowers, W., & Corenthal, C. (1985). A structured interview for the DSM-III personality disorders. *Archives of General Psychiatry*, 42, 591–596.
- Strober, M., Green, J., & Carlson, G. (1981). Reliability of psychiatric diagnosis in hospitalized adolescents: inter-rater agreement using DSM-III. *Archives of General Psychiatry*, 38, 141–145.
- Sylvester, C., Hyde, T., & Reichler, R. (1987). The Diagnostic Interview for Children and Personality Interview for Children in studies of children at risk for anxiety disorders or depression. *Journal of the American Academy of Child and Adolescent Psychiatry*, 26, 668–675.
- Tress, K. H., Bellenis, C., Brownlow, J. M., Livingston, G., & Leff, J. P. (1987). The Present State Examination change rating scale. *British Journal of Psychiatry*, 150, 201–207.
- Verhulst, F. C., Althaus, M., & Berden, G. F. M. G. (1987). The Child Assessment Schedule: Parent–child agreement and validity measures. *Journal of Child Psychology and Psychiatry*, 28, 455–466.
- Ward, C. H., Beck, A. T., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1962). The psychiatric nomenclature: Reasons for diagnostic disagreement. *Archives of General Psychiatry*, 7, 198–205.
- Welner, Z., Reich, W., Herjanic, B., Jung, K. G., & Amado, H. (1987). Reliability, validity, and parent–child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child and Adolescent Psychiatry*, 26, 649–653.
- Widiger, T. A., Mangine, S., Corbitt, E. M., Ellis, C. G., & Thomas, G. V. (1995). *Personality Disorder Interview-IV: A semistructured interview for the assessment of personality disorders*. Odessa, FL: Psychological Assessment Resources.
- Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, Jr., H. G., Rounsaville, B., & Wittchen, H.-U. (1992). The Structured Clinical Interview for DSM-III-R (SCID). II: Multisite test–retest reliability. *Archives of General Psychiatry*, 49, 630–636.
- Wing, J. K. (1983). Use and misuse of the PSE. *British Journal of Psychiatry*, 143, 111–117.
- Wing, J. K., Babor, T., Brugha, T., Burke, J., Cooper, J. E., Giel, R., Jablenski, A., Regier, D., & Sartorius, N. (1990). SCAN: Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry*, 47, 589–593.
- Wing, J. K., Birley, J. L. T., Cooper, J. E., Graham, P., & Isaacs, A. (1967). Reliability of a procedure for measuring and classifying present psychiatric state. *British Journal of Psychiatry*, 113, 499–515.
- Wing, J. K., Cooper, J. E., & Sartorius, N. (1974). *The measurement and classification of psychiatric symptoms*. London: Cambridge University Press.
- Wing, J. K., Nixon, J. M., Mann, S. A., & Leff, J. P. (1977). Reliability of the PSE (ninth edition) used in a population study. *Psychological Medicine*, 7, 505–516.
- Wittchen, H.-U. (1994). Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): A critical review. *Journal of Psychiatry Research*, 28, 57–84.
- Wittchen, H.-U., Robins, L. N., Cottler, L. B., Sartorius, N., Burke, J. D., Regier, D., & participants in the multicentre WHO/ADAMHA field trials (1991). Cross-cultural feasibility, reliability and sources of variance of the Composite International Diagnostic Interview (CIDI). *British Journal of Psychiatry*, 159, 645–653.
- Wittchen, H. -U., Semler, G., & von Zerssen, D. (1985). A comparison of two diagnostic methods: Clinical ICD

- diagnoses versus DSM-III and Research Diagnostic Criteria using the Diagnostic Interview Schedule (Version 2). *Archives of General Psychiatry*, 42, 677–684.
- World Health Organization (1973). *The international pilot study of schizophrenia, Vol. 1.*: Geneva: Author.
- World Health Organization (1990). *Composite International Diagnostic Interview (CIDI): a) CIDI-interview (version 1.0), b) CIDI-user manual, c) CIDI-training manual, d) CIDI-computer programs.* Geneva: Author.
- Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry*, 51, 225–245.
- Zimmerman, M., Pfohl, B., Stangl, D., & Corenthal, C. (1986). Assessment of DSM-III personality disorders: The importance of interviewing an informant. *Journal of Clinical Psychiatry*, 47, 261–263.

4.06

Principles and Practices of Behavioral Assessment with Children

THOMAS H. OLLENDICK

Virginia Tech, Blacksburg, VA, USA

and

ROSS W. GREENE

Harvard Medical School, Boston, MA, USA

4.06.1 INTRODUCTION	132
4.06.2 HISTORY AND DEVELOPMENT	133
4.06.3 THEORETICAL UNDERPINNINGS	134
4.06.4 DESCRIPTION OF ASSESSMENT PROCEDURES	136
4.06.4.1 Behavioral Interviews	136
4.06.4.2 Ratings and Checklists	138
4.06.4.3 Self-report Instruments	140
4.06.4.4 Self-monitoring	142
4.06.4.5 Behavioral Observation	142
4.06.5 RESEARCH FINDINGS	144
4.06.5.1 Behavioral Interviews	145
4.06.5.2 Ratings and Checklists	145
4.06.5.3 Self-report Instruments	146
4.06.5.4 Self-monitoring	146
4.06.5.5 Behavioral Observation	147
4.06.6 FUTURE DIRECTIONS	148
4.06.6.1 Developmental Factors	148
4.06.6.2 The Utility of the Multimethod Approach at Different Age Levels	149
4.06.6.3 Cultural Sensitivity	149
4.06.6.4 Measures of Cognitive and Affective Processes	150
4.06.6.5 The Role of the Child	150
4.06.6.6 Ethical Guidelines	151
4.06.7 SUMMARY	151
4.06.8 REFERENCES	151

4.06.1 INTRODUCTION

While treatment strategies derived from behavioral principles have a long and rich tradition in clinical child psychology (e.g., Holmes, 1936; Jones, 1924; Watson & Rayner, 1920), assessment practices based on these same principles have lagged, especially in the area of child behavioral assessment. In fact, many child behavioral assessment practices have been adopted, sometimes indiscriminately, from those used with adults. This practice is of dubious merit and, as we have argued elsewhere (Ollendick & Greene, 1990), it has frequently led to imprecise findings and questionable conclusions. As a result, greater attention has been focused on the development of behavioral assessment practices for children in recent years (e.g., Mash & Terdal, 1981, 1989; Ollendick & Hersen, 1984, 1993; Prinz, 1986).

As first suggested by Mash and Terdal (1981) and elaborated by Ollendick and Hersen (1984, 1993), child behavioral assessment can be defined as an ongoing, exploratory, hypothesis-testing process in which a range of specific procedures is used in order to understand a given child, group, or social ecology, and to formulate and evaluate specific intervention techniques. As such, child behavioral assessment is a dynamic, self-correcting process. It seeks to obtain information from a variety of sources in order that we might understand diverse child behavioral problems in their rich and varied contexts, and plan and evaluate behavioral interventions based upon the information obtained. Thus, assessment from this perspective is fluid (i.e., responsive to feedback and open to change(s) based on information obtained throughout the process), and it is linked intimately with treatment (i.e., assessment serves treatment). Moreover, child behavioral assessment entails more than the identification of discrete target behaviors and their controlling variables. While the importance of direct observation of target behaviors in simulated and natural settings should not be underestimated, recent advances in child behavioral assessment have incorporated a range of assessment procedures, including behavioral interviews, self-reports, ratings by significant others, and self-monitoring in addition to behavioral observations. An approach combining these procedures can best be described as a multimethod one in which an attempt is made to obtain a complete picture of the child and his or her presenting problems. Such a picture is intended to be useful in the understanding and modification of specific child behavior problems (Ollendick & Cerny, 1981; Ollendick & Hersen, 1984, 1993).

Two other primary features characterize child behavioral assessment procedures (Ollendick & Hersen, 1984, 1993). First, they must be sensitive to development, and second, they must be validated empirically. As noted by Lerner (1986, p. 41), the concept of development implies "systematic and successive changes over time in an organism." Descriptors such as "systematic" and "successive" suggest that these changes are, for the most part, orderly and that changes seen at one point in time will be influenced, at least in part, by changes that occurred at an earlier point in time. Thus development is not random nor, for that matter, discontinuous. Changes that occur at an early age (whether due to learning, an unfolding of basically predetermined structures, or some complex, interactive process) have a direct impact on subsequent development. Changes associated with development, however, create problems in selecting appropriate methods of assessment, as well as in identifying specific target behaviors for change (Ollendick & King, 1991). Behavioral interviews, self-reports, other-reports, self-monitoring, and behavioral observation may all be affected by these rapidly changing developmental processes. Further, due to "systematic and successive" change, some of these procedures may be more useful at one age than another. For example, interviews may be more difficult to conduct and self-reports less reliable with younger children, whereas self-monitoring and behavioral observations may be more reactive at older ages (Ollendick & Hersen, 1984). Age-related constraints are numerous and must be taken into consideration when selecting specific methods of assessment.

Just as child behavioral assessment procedures must be developmentally sensitive, they must also be validated empirically. All too frequently, professionals working with children have used assessment methods of convenience without sufficient regard for their psychometric properties, including their reliability, validity, and clinical utility (i.e., the degree to which assessment strategies contribute to beneficial treatment outcomes; see Hayes, Nelson, & Jarrett, 1987, for a discussion of treatment utility). Although child behavior assessors have fared somewhat better in this regard, they too have tended to design and use idiosyncratic, "convenient" tools for assessment. As we have suggested elsewhere (Ollendick & Hersen, 1984), comparison across studies is made difficult, if not impossible, and the advancement of an assessment science and technology, let alone an understanding of child behavior disorders and their effective treatments, is compromised with such an idiosyncratic approach.

While a multimethod approach that is based on developmentally sensitive and empirically validated procedures is recommended, it should be clear that a "test battery" approach is not being espoused. The specific devices to be selected depend on a host of factors, including the child's age, the nature of the referral question, the contexts in which the problematic behavior occurs, and the personnel, time, and resources available (Ollendick & Cerny, 1981). Nonetheless, given inherent limitations in each of the various procedures, as well as the desirability of obtaining as complete a picture of the child as possible, we recommend multimethod assessment whenever possible. Any one procedure, including direct behavioral observation, is not sufficient to provide a composite view of the child. The multimethod approach, if implemented, is not only helpful in assessing specific target behaviors and in determining response to behavior change, but also in understanding child behavior disorders and advancing assessment as a scientific endeavor. Based on these considerations, we offer the following tentative conclusions regarding child behavioral assessment:

(i) Children are a special and unique population. The automatic extension of adult behavioral assessment methods to children is not warranted and is often inappropriate. Further, not all "children" are alike. Clearly, a 16-year-old adolescent differs from a 12-year-old preadolescent who in turn differs from an 8-year-old middle-age child and a young 4-year-old child. Age-related variables affect the choice of methods as well as the procedures employed.

(ii) Given rapid developmental change observed in children as they grow, normative comparisons are required to ensure that appropriate target behaviors are selected and that change in behavior is related to treatment effects, and not normal developmental processes. Such comparisons require identification of suitable reference groups and information about the "natural course" of diverse child behavior problems (Ollendick & King, 1994).

(iii) Thorough child behavioral assessment involves multiple targets of change, including overt behavior, affective states, and cognitive processes. Further, such assessment entails determining the context (e.g., familial, social, cultural) in which the child's behavior occurs and the function(s) the target behaviors serve.

(iv) Given the wide range of targets for change and the imprecision of extant measures, multimethod assessment is desirable. Multimethod assessment should not be viewed simply as a test battery approach; rather, methods should be selected on the basis of their appropriateness to

the referral question. Regardless of the measures used, they should be developmentally sensitive and empirically validated.

4.06.2 HISTORY AND DEVELOPMENT

As indicated above, assessment of children's behavior problems requires a multimethod approach in which data are gathered from clinical interviews and self- and other-report sources as well as from direct behavioral observations. In this manner, important information from the cognitive and affective modalities can be obtained and integrated with behavioral data to provide a more complete picture of the child. In addition, the multimethod approach provides the clinician with necessary detail regarding perceptions and reactions of significant others in the child's life (e.g., parents, teachers, peers). It should be noted, however, that this comprehensive and inclusive assessment approach is of relatively recent origin.

In its earliest stages, behavioral assessment of children relied almost exclusively on identification and specification of discrete and highly observable target behaviors (cf. Ullmann & Krasner, 1965). As such, assessment was limited to gathering information solely from the motoric (i.e., behavioral) response modality. This early assessment approach followed logically from theoretical assumptions of the operant school of thought which was in vogue at the time. Early on, behaviorally oriented clinicians posited that the only appropriate behavioral domain for empirical study was that which was directly observable (Skinner, 1953). Contending that objective demonstration of behavior change following intervention was of utmost importance, behaviorists relied upon data that could be measured objectively. Subjectivity, and the inferential process associated with it, were eschewed. Hence the frequency, intensity, and duration of problematic behaviors (i.e., "hard core" measures) were pursued. Although existence of cognitions and affective states was not denied, they were not deemed appropriate subject matter for experimental analysis.

As behavioral treatment approaches with children were broadened to include cognitive and self-control techniques in the 1970s (e.g., Bandura, 1977; Kanfer & Phillips, 1970; Kendall & Hollon, 1980; Meichenbaum, 1977), it became apparent that assessment strategies would have to expand into the cognitive and affective domains as well. Furthermore, even though operant techniques were shown to be efficacious in producing behavior change under controlled conditions, the clinical significance

of these changes was less evident. The issue of clinical significance of behavior change is especially important in child behavioral assessment because children are invariably referred for treatment by others (e.g., parents, teachers). Once treatment goals have been identified, the ultimate index of treatment efficacy lies in the referral source's perceptions of change. Hence, other-report measures become as important as direct observational ones.

Furthermore, the scope of behavioral assessment has been expanded to include the impact of large-scale social systems (e.g., schools, neighborhoods) on the child's behavior (e.g., Patterson, 1976; Wahler, 1976). Although inclusion of these additional factors serves to complicate the assessment process, they are an indispensable part of modern-day child behavioral assessment. The ideologies and expectations of seemingly distal social systems often have immediate and profound effects on individual behavior (see Winett, Riley, King, & Altman, 1989, for discussion of these issues).

In sum, child behavioral assessment has progressed from sole reliance on measurement of target behaviors to a broader approach that takes into account cognitive and affective processes of the child that serve to mediate behavior change. Further, the social contexts (i.e., families, schools, communities) in which the problematic behaviors occur have been targeted for change. The assessment techniques that accompany this approach include behavioral interviews and self- and other-report instruments. These measures are utilized in addition to direct behavioral observation which remains the cornerstone of behavioral assessment (Mash & Terdal, 1981, 1989; Ollendick & Hersen, 1984, 1993).

4.06.3 THEORETICAL UNDERPINNINGS

Although behaviorism has had an historical development of its own, it is safe to state that the increased popularity of the behavioral approach has flourished, at least in part, due to dissatisfaction with the psychodynamic approach. A reflection of this dissatisfaction is that virtually all discussions of behavioral assessment are carried out through comparison and contrast with traditional assessment approaches (e.g., Bornstein, Bornstein, & Dawson, 1984; Cone & Hawkins, 1977; Goldfried & Kent, 1972; Hayes, Nelson, & Jarrett, 1986; Mash & Terdal, 1981, 1989; Mischel, 1968; Ollendick & Hersen, 1984, 1993). Though such comparisons often result in oversimplification of both approaches, they serve to elucidate theoretical underpinnings of the behavioral

approach and its unique contributions. In this section, we will contrast the theoretical assumptions that guide behavioral and traditional assessment and discuss the practical implications of these assumptions for child behavioral assessment.

The most fundamental difference between traditional and behavioral assessment lies in the conception of "personality" and behavior (we place the construct "personality" in quotations because early behaviorists would have objected to use of this term, given its subjectivity and imprecise meaning). In the traditional assessment approach, personality is viewed as a reflection of underlying and enduring traits, and behavior is assumed to be caused by these internal personality characteristics ("personalism"). Aggressive behavior, for example, is assumed to reside "in" the child and to be caused by an underlying dynamic process attributed, perhaps, to hostility or anger and resulting from deep-seated intrapsychic conflict. "Aggression," it is said, is caused by the underlying hostility/anger. In contrast, behavioral approaches have avoided references to underlying personality constructs, focusing instead on what the child does under specific conditions. From the behavioral perspective, "personality" refers to patterns rather than causes of behavior (Staats, 1975, 1986). Furthermore, behavior is viewed as a result of current environmental factors ("situationalism") and of current environmental factors interacting with organismic variables ("interactionism"). Thus the role of the current environment is stressed more in behavioral assessment than in traditional assessment. The focus of assessment is on what the child does in that situation rather than on what the child has or "is" (Mischel, 1968). As a result, a lower level of inference is required in behavioral assessment than in traditional assessment.

It is important not to oversimplify the behavioral view of the causes of behavior, however. It has often been erroneously asserted that the behavioral approach focuses on external determinants of behavior to the exclusion of organismic states or internal cognitions and affects. To be sure, behavioral views of childhood disorders have emphasized the significant role of environmental factors in the manifestation and maintenance of behavior. However, organismic variables that influence behavior are not ignored or discounted. Among the organismic variables—dubbed cognitive social learning person variables (CSLPVs) by Mischel (1973)—that have been found to be important are competencies (skills which children possess such as social skills, problem-solving skills), encoding strategies (the manner

in which children perceive or encode information about their environment), expectancies (expectancies about performance, including self-efficacy and outcome expectancies), subjective values (children's likes or dislikes, preferences or aversions), and self-regulatory systems and plans (children's capacity for and manner of self-imposing goals and standards and self-administering consequences for their behavior). A wide array of self-report instruments tapping CSLPVs and related cognitive and affective modalities for use in child behavioral assessment have been reviewed recently by us (Greene & Ollendick, *in press*). A thorough behavioral assessment should attempt to identify controlling variables, whether environmental or organismic in nature. As Mash and Terdal (1981) point out, "the relative importance of organismic and environmental variables and their interaction . . . should follow from a careful analysis of the problem" (p. 23).

The traditional conception of personality as made up of stable and enduring traits implies that behavior will be relatively persistent over time and consistent across situations. The behavioral view, in contrast, has been one of situational specificity; that is, because behavior is in large part a function of situational determinants and CSLPVs that are enacted only under specified conditions, a child's behavior will change as these situational factors are altered or the person variables are engaged. Similarly, consistency of behavior across the temporal dimension is not necessarily expected. Hence, as noted above, an aggressive act such as a child hitting another child would be seen from the traditional viewpoint as a reflection of underlying hostility which, in turn, would be hypothesized to be related to early life experiences or intrapsychic conflict. Little or no attention would be given to specific situational factors or the environmental context in which the aggressive act occurred. From the behavioral perspective, an attempt is made to identify those variables that elicit and maintain the aggressive act in that particular situation. That the child may aggress in a variety of situations is explained in terms of his or her learning history in which reinforcing consequences have been obtained for past aggressive acts (which help shape CSLPVs), and not in terms of an underlying personality trait of hostility. From this analysis, it is clear that actual behavior is of utmost importance to behaviorists, because it represents a sample of the child's behavioral repertoire in a specific situation. From the traditional viewpoint, the behavior assumes importance only insofar as it is a sign of some underlying trait.

These differing assumptions have implications for the assessment process. In behavioral assessment, the emphasis on situational specificity necessitates an assessment approach that samples behavior across a number of settings. Hence assessment of the child's behavior at home, in school, and on the playground is important in addition to information obtained in the clinic setting. Furthermore, it is not assumed that information obtained from these various settings will be consistent. The child may behave aggressively in school and on the playground with peers but not at home with siblings or parents. Or conversely, the child might behave aggressively at home but not at school or when with his or her peers. This lack of consistency in behavior would be problematic for the traditional approach, but not for the behavioral approach. Similarly, the notion of temporal instability requires the child's behavior be assessed at several points in time from a behavioral perspective, whereas such measurements across time would be less critical for the traditional approach.

At one point in time, it was relatively easy to differentiate behavioral from traditional assessment on the basis of the methods employed. Direct behavioral observation was the defining characteristic and often the sole assessment technique of the behavioral approach, whereas clinical interviews, self-report measures, and projective techniques characterized traditional assessment. However, as behavioral assessment was expanded to include a wider repertoire of assessment methods, differentiating behavioral and traditional assessments simply on the basis of assessment methods used has become more difficult. It is not uncommon for behaviorists to utilize information from clinical interviews and self-report instruments, and to pursue perceptions and expectancies of significant others in the child's environment. Thus there is considerable overlap in actual assessment practices, with one notable exception. Rarely, if ever, would projective techniques be utilized by the child behavioral assessor. The primary difference between traditional and behavioral assessment lies then not in the methods employed, but rather in the manner in which data from assessment sources are utilized. Traditional approaches interpret assessment data as signs of underlying personality functioning. These data are used to diagnose and classify the child and to make prognostic statements. From the behavioral perspective, assessment data are used to identify target behaviors and their controlling conditions (again, be they overt or covert). Information obtained from assessment serves as a sample of the child's behavior under specific circumstances. This information guides the

selection of appropriate treatment procedures. Because behavioral assessment is ongoing, such information serves as an index by which to evaluate critically the effects of treatment and to make appropriate revisions in treatment. Further, because assessment data are viewed as samples of behavior, the level of inference is low, whereas a high level of inference is required when one attempts to make statements about personality functioning from responses to interview questions or test items.

In addition to these differences, Cone (1986) has highlighted the nomothetic and ideographic distinction between traditional and behavioral assessment. Stated briefly, the nomothetic approach is concerned with the discovery of general laws as they are applied to large numbers of children. Usually, these laws provide heuristic guidelines as to how certain variables are related to one another. Such an approach can be said to be variable-centered because it deals with particular characteristics (traits) such as intelligence, achievement, assertion, aggression, and so on. In contrast, the ideographic approach is concerned more with the uniqueness of a given child and is said to be child-centered rather than variable-centered. Unlike the nomothetic approach, the ideographic perspective emphasizes discovery of relationships among variables uniquely patterned in each child. The ideographic approach is most akin to the behavioral perspective, whereas the nomothetic approach is closely related to the traditional approach. As Mischel (1968) observed, "Behavioral assessment involves an exploration of the unique or idiosyncratic aspects of the single case, perhaps to a greater extent than any other approach" (p. 190). Cone (1986) illustrates how the ideographic/nomothetic distinction relates to the general activities of behavioral assessors by exploring five basic questions: What is the purpose of assessment? What is its specific subject matter? What general scientific approach guides this effort? How are differences accounted for? And, to what extent are currently operative environmental variables considered? Although further discussion of these important issues is beyond the scope of the present chapter, Cone's schema helps us recognize the pluralistic nature of behavioral assessment and calls our attention to meaningful differences in the practices contained therein. As Cone (1986) concludes, "There is not one behavioral assessment, there are many" (p. 126). We agree.

In sum, traditional and behavioral assessment approaches operate under different assumptions regarding the child's behavior. These assumptions, in turn, have implications for

the assessment process. Of paramount importance for child behavior assessors is the necessity of tailoring the assessment approach to the specific difficulties of the child in order to identify the problem accurately, specify treatment, and evaluate treatment outcome. Such tailoring requires ongoing assessment from a number of sources under appropriately diverse stimulus conditions.

4.06.4 DESCRIPTION OF ASSESSMENT PROCEDURES

Multimethod behavioral assessment of children entails use of a wide variety of specific procedures. As behavioral approaches with children evolved from sole reliance on operant procedures to those involving cognitive and self-control procedures, the methods of assessment have changed accordingly. Identification of discrete target behaviors has been expanded to include assessment of cognitions and affects, as well as large-scale social systems that affect the child (e.g., families, schools, communities).

Information regarding these additional areas can be obtained most efficiently through behavioral interviews, self-reports, and other-reports. Cone (1978) has described these assessment methods as indirect ones; that is, while they may be used to measure behaviors of clinical relevance, they are obtained at a time and place different from when the behaviors actually occurred. In both behavioral interviews and self-report questionnaires, a verbal representation of the behaviors of interest is obtained. Other-reports, or ratings by others such as parents or teachers, are also included in the indirect category because they involve retrospective descriptions of behavior. Generally, a significant person in the child's environment (e.g., at home or school) is asked to rate the child based on previous observations in that setting (recollections).

As noted by Cone (1978), ratings such as these should not be confused with direct observation methods, which assess behaviors of interest at the time and place of their occurrence. Of course, information regarding cognition and affects, as well as the situations or settings in which they occur, can also be obtained through direct behavioral observations, either by self-monitoring or through trained observers. In the sections that follow, both indirect and direct methods are reviewed.

4.06.4.1 Behavioral Interviews

The first method of indirect assessment to be considered is the behavioral interview. Of the

many procedures employed by behavioral clinicians, the interview is the most widely used (Swann & MacDonald, 1978) and is generally considered an indispensable part of assessment (Gross, 1984; Linehan, 1977). Behavioral interviews are frequently structured to obtain information about the target behaviors and their controlling variables and to begin the formulation of specific treatment plans. While the primary purpose of the behavioral interview is to obtain information, we have found that traditional "helping" skills including reflections, clarifications, and summary statements help put children and their families at ease and greatly facilitate collection of this information (Ollendick & Cerny, 1981). As with traditional therapies, it is important to establish rapport with the child and family and to develop a therapeutic alliance (i.e., agreement on the goals and procedures of therapy) in the assessment phase of treatment (Ollendick & Ollendick, 1997).

Undoubtedly, the popularity of the behavioral interview is derived in part from practical considerations associated with its use. While direct observation of target behaviors remains the hallmark of behavioral assessment, such observations are not always practical or feasible. At times, especially in outpatient therapy in clinical settings, the clinician might have to rely on children's self-report as well as that of their parents to obtain critical detail about problem behaviors and their controlling variables. Further, the interview affords the clinician the opportunity to obtain information regarding overall functioning in a number of global areas (e.g., home, school, neighborhood), in addition to specific information about particular problem areas. The flexibility inherent in the interview also allows the clinician to build a relationship with the child and the family and to obtain information that might otherwise not be revealed. As noted early on by Linehan (1977), some family members may be more likely to divulge information verbally in the context of a professional relationship than to write it down on a form to be entered into a permanent file. In our experience, this is not an uncommon occurrence. That is, certain family members report little or no difficulties on intake reports or on self-report measures, yet they divulge a number of problem areas during the structured behavioral interview.

In addition, the interview allows the clinician the opportunity to observe the family as a whole and to obtain information about the familial context in which the problem behaviors occur. Several interrelated issues may arise when child behavioral assessment is expanded to include the family unit (Evans & Nelson, 1977;

Ollendick & Cerny, 1981). First, children rarely refer themselves for treatment; invariably, they are referred by adults whose perceptions of problems may not coincide with the child's view. This is especially true when problems are centered around externalizing behaviors such as oppositional or disruptive behaviors, less so with internalizing behaviors (e.g., anxiety or depression). Moreover, it is not uncommon for the perception of one adult to differ from that of another (i.e., the mother and father disagree, or the teacher and the parents disagree; cf. Achenbach, McConaughy, & Howell, 1987). A second issue, related to the first, is the determination of when child behaviors are problematic and when they are not. Normative developmental comparisons are useful in this regard (Lease & Ollendick, 1993; Ollendick & King, 1991). It is not uncommon for parents to refer 3-year-olds who wet the bed, 5-year-olds who reverse letters, 10-year-olds who express interest in sex, and 13-year-olds who are concerned about their physical appearance. Frequently, these referrals are based on parental uneasiness or unrealistic expectations rather than genuine problems (see Campbell, 1989, for further discussion of these issues). Finally, problematic family interactions (especially parent-child interactions) are frequently observed in families in which a particular child has been identified and referred for treatment (cf. Dadds, Rapee, & Barrett, 1994; Patterson, 1976, 1982). These interactions may not be a part of the parents' original perception of the "problem." Furthermore, assessment of such interactions allows the clinician an opportunity to observe the verbal and nonverbal behaviors of the family unit in response to a variety of topics, and of family members in response to each other. Structured interviews assessing parent-child interactions have been developed for a number of behavior problems (e.g., Barkley, 1987; Dadds et al., 1994).

Ideally, evaluation of parental perceptions and parent-child interactions will enable the clinician to conceptualize the problematic behaviors and formulate treatment plans from a more comprehensive, integrated perspective. However, the above discussion is not meant to imply that the behavioral interview should be limited to the family; in many instances, the practices described above should be extended to adults outside the family unit, such as teachers, principals, and physicians, and to environments beyond the home, including schools and day-care centers. For example, if a problem behavior is reported to occur primarily at school, assessing the perceptions and behavioral goals of a teacher and/or principal will be necessary (Greene, 1995, 1996), and evaluating teacher-

student interactions may prove more productive than observing parent-child interactions during the clinical interview. Finally, the clinician should approach the behavioral interview with caution and avoid blind acceptance of the premise that a "problem" exists "in" the child. Information obtained in a comprehensive assessment may reveal the behavior of the identified client is only a component of a more complex clinical picture involving parents, siblings, other adults, and/or social systems.

In sum, an attempt is made during the behavioral interview to obtain as much information as possible about the child, his or her family, and other important individuals and environments. While the interview is focused around specific target behaviors, adult-child interactions and adult perceptions of the problem may also be assessed. These perceptions should be considered tentative, however, and used primarily to formulate hypotheses about target behaviors and their controlling variables and to select additional assessment methods to explore target behaviors in more depth (e.g., rating scales, self-reports, self-monitoring, and behavioral observations). The behavioral interview is only the first step in the assessment process.

Brief mention should also be made here of structured diagnostic interviews and their role in child behavioral assessment. In some instances, most notably when a diagnosis is required, it may be desirable for the clinician to conduct a structured diagnostic interview. In general, diagnostic interviews are oriented toward obtaining specific information to determine if a child "meets" diagnostic criteria for one or more specific diagnoses included in the *Diagnostic and statistical manual of mental disorders* (4th ed., *DSM-IV*) (American Psychiatric Association, 1994) or the *International classification of diseases* (10th ed., *ICD-10*; World Health Organization, 1991). Such interviews facilitate collection of data relative to a broad range of "symptoms" (i.e., behaviors) and psychiatric diagnoses. Several "omnibus" diagnostic interviews are available, including the Diagnostic Interview Schedule for Children-Version 2.3 (Shaffer, 1992), which was recently revised to reflect *DSM-IV* criteria. Other diagnostic interviews are oriented toward a specific domain such as anxiety (e.g., the Anxiety Disorders Interview Schedule for Children; Silverman & Nelles, 1988). It, too, has recently been revised to incorporate *DSM-IV* criteria (Silverman & Albano, 1996). Both child and parent forms of these interviews are available. Although these structured diagnostic interviews provide a wealth of information, they are limited by an overemphasis on diagnostic

categories (to the exclusion of important details regarding specific target behaviors and their controlling variables), weak or untested reliability for children under age 11, low correspondence between responses of children and their parents, and categorical vs. dimensional scoring criteria (McConaughy, 1996). Further, structured diagnostic interviews often do not yield specific information about contextual factors associated with the child's problematic behavior; thus, when a diagnostic interview is used, it needs to be supplemented with a problem-focused interview. In our opinion, diagnostic interviews should not be considered as replacements for problem-focused interviews; rather they should be viewed as complementary.

4.06.4.2 Ratings and Checklists

Following the initial behavioral interview and, if necessary, the diagnostic interview, significant others in the child's environment may be requested to complete rating forms or checklists. In general, these forms are useful in providing an overall description of the child's behavior, in specifying dimensions or response clusters that characterize the child's behavior, and in serving as outcome measures for the evaluation of treatment efficacy. Many of these forms contain items related to broad areas of functioning such as school achievement, peer relationships, activity level, and self-control. As such, they provide a cost-effective picture of children and their overall level of functioning. Further, the forms are useful in eliciting information that may have been missed in the behavioral interview (Novick, Rosenfeld, Bloch, & Dawson 1966). Finally, the forms might prove useful in the search for the best match between various treatments (e.g., systematic desensitization, cognitive restructuring, and self-control) and subtypes of children as revealed on these forms (Ciminero & Drabman, 1977).

The popularity of omnibus rating forms and checklists is evident in the number of forms currently available (McMahon, 1984). Three of the more frequently used forms are described here: the Behavior Problem Checklist (Quay & Peterson, 1967, 1975) and its revision (Quay & Peterson, 1983); the Child Behavior Checklist (Achenbach, 1991a, 1991b); and the recently developed Behavior Assessment System for Children (Reynolds & Kamphaus, 1992). Based on Peterson's (1961) early efforts to sample diverse child behavior problems, the Revised Behavior Problem Checklist consists of 89 items, each rated on a three-point severity scale. While some of the items are general and require considerable inference (e.g., lacks self-

confidence, jealous), others are more specific (e.g., cries, sucks thumb). Six primary dimensions or response clusters of child behavior have been identified on this scale: conduct problems, socialized aggression, attention problems, anxiety-withdrawal, psychotic behavior, and motor excess. Interestingly, the two primary problem clusters found on this checklist are similar to those found in numerous factor-analytic studies of other rating forms and checklists. These two factors or response clusters represent consistent dimensions of child behavior problems, reflecting externalizing (e.g., acting out) and internalizing (e.g., anxiety, withdrawal) dimensions of behavior (Achenbach, 1966).

While the Behavior Problem Checklist has a rather lengthy history and is one of the most researched scales, it does not include the rating of positive behaviors and, hence, does not provide a basis on which to evaluate more appropriate, adaptive behaviors. A scale that does assess appropriate behaviors, as well as inappropriate ones, is the Child Behavior Checklist (CBCL; Achenbach, 1991a, 1991b; Achenbach & Edelbrock, 1989). The scale, designed for both parents and teachers, contains both social competency and behavior problem items. The parent-completed CBCL is available in two formats depending on the age of the child being evaluated (i.e., 2–3 years and 4–18 years). The CBCL 4–18, for example, consists of 112 items rated on a three-point scale. Scored items can be clustered into three factor-analyzed profiles: social competence, adaptive functioning, and syndrome scales. The latter includes eight syndrome scales: withdrawn, somatic complaints, anxious/depressed, social problems, thought problems, attention problems, aggressive behavior, and delinquent behavior. Social competency items examine the child's participation in various activities (e.g., sports, chores, hobbies) and social organizations (e.g., clubs, groups), as well as performance in the school setting (e.g., grades, placements, promotions). The teacher-completed CBCL (TRF; Teacher Report Form) also consists of 112 items which are fairly similar, but not completely identical to, those found on the CBCL completed by parents. The scored items from the teacher form cluster into the same three factor-analyzed profiles; further, the eight syndrome scales are the same for the two measures, allowing for cross-informant comparisons. As with Quay and Peterson's Behavior Problem Checklist, some of the items are general and require considerable inference (e.g., feels worthless, acts too young, fears own impulses), while others are more specific and easily scored (e.g., wets bed, sets fires, destroys own things). Broad-

band grouping of the factors reflects the aforementioned internalizing and externalizing behavioral dimensions.

Although the Behavior Problem Checklist and Child Behavior Checklist have enjoyed considerable success, the recently developed Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) represents a challenge to both of these well-established rating scales. Like these other instruments, the BASC is an omnibus checklist composed of parent, teacher, and child versions. It also contains a developmental history form and a classroom observation form. Most similar to Achenbach's Child Behavior Checklist (Achenbach, 1991a), Teacher's Report Form (Achenbach, 1991b), and Youth Self-Report (Achenbach, 1991c), the parent, teacher, and self-report forms of the BASC contain items that tap multiple emotional and behavioral domains and produce scale scores that represent pathological and adaptive characteristics of the child. Unlike the empirically derived scales of Achenbach's checklists, however, the scales of the BASC were created conceptually to represent content areas relevant to assessment and classification in clinical and educational settings.

For example, the BASC Parent Rating Scale (BASC-PRS) yields *T* scores in broad externalizing and internalizing domains as well as in specific content areas, including aggression, hyperactivity, conduct problems, attention problems, depression, anxiety, withdrawal, somatization, and social skills.

In addition, it provides *T* scores in areas of social competency such as leadership and adaptability. Preschool (ages 4–5), child (ages 6–11), and adolescent (ages 12–18) forms are available. Recent findings suggest the utility of this instrument with both clinical and educational populations and in identifying youth at risk for maladaptive outcomes (cf, Doyle, Ostrander, Skare, Crosby, & August, 1997). Although initial findings associated with its use appear promising, much more research is needed before its routine acceptance can be endorsed.

In addition to these more general rating forms, rating forms specific to select problem areas are also available for use in child behavioral assessment. Three such forms have been chosen for purposes of illustration: one used in the assessment of an internalizing dimension (fears/anxiety), another in the assessment of an externalizing dimension (defiance/noncompliance), and the final one in measuring a specific area of social competency.

The Louisville Fear Survey Schedule for Children (Miller, Barrett, Hampe, & Noble,

1972) contains 81 items that address an extensive array of fears and anxieties found in children and adolescents. Each item is rated on a three-point scale by the child's parents. Responses to specific fear items can be used to subtype fearful children. For example, Miller et al. (1972) were able to differentiate among various subtypes of school-phobic children on the basis of this instrument.

The Home Situations Questionnaire (HSQ; Barkley, 1981; Barkley & Edelbrock, 1987) contains 16 items representing home situations in which noncompliant behavior may occur (e.g., while playing with other children, when asked to do chores, and when asked to do homework). For each situation, parents indicate whether noncompliant behavior is a problem and then rate each of the 16 problematic situations on a nine-point scale (mild to severe); thus the scale assesses both the number of problem situations and the severity of noncompliant behavior. The HSQ has been shown to be sensitive to stimulant-drug effects (Barkley, Karlsson, Strzelecki, & Murphy, 1984), to discriminate children with behavior problems from normal children (Barkley, 1981), and to be sensitive to the effects of parent training programs (Pollard, Ward, & Barkley, 1983). The HSQ was selected for inclusion in this chapter because it may also be used in conjunction with a companion scale, the School Situations Questionnaire (SSQ; Barkley, 1981; Barkley & Edelbrock, 1987), which is completed by teachers. This scale includes 12 school situations most likely to be problematic for clinic-referred, noncompliant children, including "during lectures to the class," "at lunch," and "on the bus." Teachers rate the occurrence and severity of noncompliant behavior on a scale identical to that of the HSQ. In earlier sections, we emphasized the importance of assessing child behavior in multiple environments; the HSQ and SSQ are representative of recent efforts to develop measures for this purpose, thus providing us important contextual information about specific problem behaviors.

In some instances, it may be useful to obtain more information about a positive, adaptive domain of behavior—such as social skills or self-regulation—than that provided for by omnibus rating scales such as the Revised Behavior Problem Checklist (Quay & Peterson, 1983) or the Child Behavior Checklist (Achenbach, 1991a, 1991b). For example, the Social Skills Rating System (Gresham & Elliot, 1990), a 55-item questionnaire, provides specific information about a child's behavior in three domains (social skills, problem behaviors, and academic competence). Parent, teacher, and

self-rating forms are available. In general, this instrument provides important and detailed information about academic and social competence that can be used to supplement information obtained from the more generic rating scales.

In sum, a variety of other-report instruments are available. As noted earlier, these forms must be considered indirect methods of assessment because they rely on retrospective descriptions of the child's behavior. For all of these scales, an informant is asked to rate the child based on past observations of that child's behavior. Global scales such as the Revised Behavior Problem Checklist, Child Behavior Checklist, and Behavior Assessment Scale for Children comprehensively sample the range of potential behavior problems, while more specific scales such as the Louisville Fear Survey Schedule for Children, the Home Situations Questionnaire, and the Social Skills Rating System provide detailed information about particular maladaptive or adaptive behaviors. Both types of other-report instruments provide useful, albeit different, information in the formulation and evaluation of treatment programs.

4.06.4.3 Self-report Instruments

Concurrent with the collection of other-reports regarding the child's behavior from significant others, self-reports of attitudes, feelings, and behaviors may be obtained directly from the child. As noted earlier, behaviorists initially eschewed such data, maintaining that the only acceptable datum was observable behavior. To a large extent, this negative bias against self-report was an outgrowth of early findings indicating that reports of subjective states did not always coincide with observable behaviors (Finch & Rogers, 1984). While congruence in responding is, in fact, not always observed, contemporary researchers have cogently argued that children's perceptions of their own behavior and its consequences may be as important for behavior change as the behavior itself (Finch, Nelson, & Moss, 1993; Ollendick & Hersen, 1984, 1993). Furthermore, as noted earlier, although different assessment procedures may yield slightly different information, data from these sources should be compared and contrasted in order to produce the best picture of the child and to derive treatment goals and procedures. Although self-report instruments have specific limitations, they can provide valuable information about children and their presenting problem; furthermore, they can be used as an index of change following treatment.

A wide array of self-report instruments have been developed for children. Some self-report instruments focus on a broad range of behavioral, cognitive, and affective functioning, as in the case of the Youth Self-report (Achenbach, 1991c). Other self-report instruments tap more specific areas of interest, such as anger (Nelson & Finch, 1978), anxiety (Reynolds & Richmond, 1985; Spielberger, 1973), assertion (Deluty, 1979; Ollendick, 1983a), depression (Kovacs, 1985), and fear (Scherer & Nakamura, 1968). Each of these instruments has been carefully developed and empirically validated. Three of the more frequently used instruments we have found to be useful in our clinical practice will be described briefly.

Spielberger's State-Trait Anxiety Inventory for Children (1973) consists of 20 items that measure state anxiety and 20 items that tap trait anxiety. The state form is used to assess transient aspects of anxiety, while the trait form is used to measure more global, generalized aspects of anxiety. Combined, the two scales can provide both process and outcome indices of change in self-reported anxiety. That is, the state form can be used to determine session-by-session changes in anxiety, while the trait form can be used as a pretreatment, post-treatment, and follow-up measure of generalized anxiety. A clear advantage of this instrument is that the state scale is designed so responses to relatively specific anxiety-producing situations can be determined. For example, children can be instructed to indicate how they feel "at this moment" about standing up in front of class, leaving home for summer camp, or being ridiculed by peers. Further, cognitive, motoric, and physiologic indicators of anxiety can be endorsed by the child (e.g., feeling upset, scared, mixed up, jittery, or nervous). Responses to items are scored on a three-point scale (e.g., "I feel very scared/scared/not scared"). Finally, the pervasiveness of the anxiety response can be measured by the trait form. The Spielberger scales are most useful for children aged 9–12, but have been used with both younger children and adolescents as well.

A second instrument that has been used frequently in child behavioral assessment is the Fear Survey Schedule for Children (Scherer & Nakamura, 1968) and its revision (Ollendick, 1983b). In the revised scale, designed to be used with younger and middle-age (9–12) children, children are instructed to rate their level of fear to each of 80 items on a three-point scale. They are asked to indicate whether a specific fear item (e.g., having to go to school, snakes, dark places, riding in a car) frightens them "not at all," "some," or "a lot." Factor analysis of the scale has revealed five primary factors: fear of

failure or criticism, fear of the unknown, fear of injury and small animals, fear of danger and death, and medical fears. This pattern of fear has been shown to be relatively invariant across several nationalities, including American (Ollendick, Matson & Hetsel, 1985), Australian (Ollendick, King, & Frary, 1989), British (Ollendick, Yule, & Ollier, 1991), Chinese (Dong, Yang, & Ollendick, 1994), and Nigerian youth (Ollendick, Yang, King, Dong, & Akande, 1996). Further, it has been shown that girls report more fear than boys in these various countries, that specific fears change developmentally, and that the most prevalent fears of boys and girls have remained unchanged over the past 30 years (although some differences have been noted across nationalities). Such information is highly useful when determining whether a child of a specific age and gender is excessively fearful. Further, the instrument has been used to differentiate subtypes of phobic youngsters whose fear of school is related to separation anxiety (e.g., death, having parents argue, being alone) from those whose fear is due to specific aspects of the school situation itself (e.g., taking a test, making a mistake, being sent to the principal). When information from this instrument is combined with that from parents on the Louisville Fear Survey Schedule for Children (Miller et al., 1972), a relatively complete picture of the child's characteristic fear pattern can be obtained.

The final self-report instrument to be reviewed is Kovacs's (1985) Children's Depression Inventory (CDI). Since the mid-1980s, no other area in clinical child psychology has received more attention than depression in children. A multitude of issues regarding its existence, nature, assessment, and treatment have been examined (Cantwell, 1983; Rutter, 1986). One of the major obstacles to systematic investigations in this area has been the absence of an acceptable self-report instrument, and the CDI appears to meet this need. The instrument is a 27-item severity measure of depression based on the well-known Beck Depression Inventory. Each of the 27 items consists of three response choices designed to range from mild depression to fairly severe and clinically significant depression. Kovacs reports that the instrument is suitable for middle-age children and adolescents (8–17 years of age). We have found the instrument to be useful with younger children as well, especially when items are read aloud and response choices are depicted on a bar graph. Smucker, Craighead, Craighead, and Green (1986) have provided additional psychometric data on the CDI. Overall, they conclude it is a reliable, valid, and clinically useful instrument for children and adolescents.

In sum, a variety of self-report instruments are available. As with other-report forms, self-reports should be used with appropriate caution and due regard for their specific limitations. Because they generally involve the child's retrospective rating of attitudes, feelings, and behaviors, they too must be considered indirect methods of assessment (Cone, 1978). Nevertheless, they can provide valuable information regarding children's own perception of their behavior.

4.06.4.4 Self-monitoring

Self-monitoring differs from self-report in that it constitutes an observation of clinically relevant target behaviors (e.g., thoughts, feelings, actions) at the time of their occurrence (Cone, 1978). As such, it is a direct method of assessment. Self-monitoring requires children to observe their own behavior and then to record its occurrence systematically. Typically, the child is asked to keep a diary, place marks on a card, or push the plunger on a counter as the behavior occurs or shortly thereafter. Although self-monitoring procedures have been used with both children and adults, at least three considerations must be attended to when such procedures are used with younger children (Shapiro, 1984): behaviors should be clearly defined, prompts to use the procedures should be readily available, and rewards for their use should be provided. Some younger children will be less aware of when the target behavior is occurring and will require coaching and assistance prior to establishing a monitoring system. Other young children may have difficulty remembering exactly what behaviors to monitor and how those behaviors are defined. For these reasons, it is generally considered advisable to provide the child with a brief description of the target behavior or, better yet, a picture of it, and to have the child record only one or two behaviors at a time. In an exceptionally sensitive application of these guidelines, Kunzelman (1970) recommended the use of COUNTOONS, simple stick figure drawings that depict specific behaviors to be self-monitored. Children are instructed to place a tally mark next to the picture when the behavior occurs. For example, a girl monitoring hitting her younger brother may be given an index card with a drawing of a girl hitting a younger boy and instructed to mark each time she does what the girl in the picture is doing. Of course, in a well-designed program, the girl might also be provided with a picture of a girl and a younger boy sharing toys and asked as well to mark each time she emits the appropriate behavior. Such pictorial cues

serve as visual prompts for self-monitoring. Finally, children should be reinforced profusely following successful use of self-monitoring.

In general, methods of self-monitoring are highly variable and depend on the specific behavior being monitored and its place of occurrence. For example, Shapiro, McGonigle, and Ollendick (1980) had mentally retarded and emotionally disturbed children self-monitor on-task behavior in a school setting by placing gummed stars on assignment sheets; Ollendick (1981) had children with tic disorders place tally marks upon the occurrence of tics on a colored index card carried in the child's pocket; and Ollendick (1995) had adolescents diagnosed with panic disorder and agoraphobia indicate the extent of their agoraphobic avoidance on a 1–5 scale each time they encountered the feared situation. He also had the adolescents indicate their confidence (i.e., self-efficacy) in coping with their fear on a similar 1–5 scale. In our clinical work, we have also used wrist counters with children whose targeted behaviors occur while they are “on the move.” Such a device is not only easy to use, but serves as a visual prompt to self-record. The key to successful self-monitoring in children is the use of recording procedures that are uncomplicated. They must be highly portable, simple, time-efficient, and relatively unobtrusive (Greene & Ollendick, in press).

In sum, self-monitoring procedures represent a direct means of obtaining information about the target behaviors as well as their antecedents and consequences. While specific monitoring methods may vary, any procedure that allows the child to monitor and record presence of the targeted behaviors can be used. When appropriate procedures are used, self-monitoring represents a direct and elegant method of assessment (Ollendick & Greene, 1990; Ollendick & Hersen, 1993).

4.06.4.5 Behavioral Observation

Direct observation of the child's behavior in the natural environment is the hallmark of child behavioral assessment. As described by Johnson and Bolstad (1973), the development of naturalistic observation procedures represents one of the major, if not the major, contributions of the behavioral approach to assessment and treatment of children. A direct sample of the child's behavior at the time and place of its occurrence is obtained with this approach. As such, it is the least inferential of the assessment methods described heretofore. However, behavioral observations in the naturalistic environment should not be viewed as better than other

methods of assessment. Rather, direct observations should be viewed as complementary to the other methods, with each providing different and valuable information.

In behavioral observation systems, a single behavior or set of behaviors that have been identified as problematic (generally through the aforementioned procedures) are operationally defined, observed, and recorded in a systematic fashion. In addition, events that precede and follow behaviors of interest are recorded and subsequently used in development of specific treatment programs. Although Jones, Reid, and Patterson (1975) have recommended use of "trained impartial observer-coders" for collection of these data, this is rarely possible in the practice of child behavioral assessment in the clinical setting. Frequently, time constraints, lack of trained personnel, and insufficient resources mitigate against the use of highly trained and impartial observers. In some cases, we have used significant others in the child's environment (e.g., parents, teachers, siblings) or the children themselves as observers of their own behavior. Although not impartial, these observers can be trained adequately to record behaviors in the natural environment. In other cases, behavioral clinicians have resorted to laboratory or analogue settings that are similar to, but not identical to, the natural environment. In these simulated settings, children may be asked to behave as if they are angry with their parents, to role play assertive responding, or to approach a highly feared object. Behaviors can be directly observed or videotaped (or audio-taped) and reviewed retrospectively. The distinguishing characteristic of behavioral observations, whether made in the naturalistic environment or in simulated settings, is that a direct sample of the child's behavior is obtained.

A wide variety of target behaviors have been examined using behavioral observation procedures. These behaviors have varied from relatively discrete behaviors like enuresis and tics, that require relatively simple and straightforward recording procedures, to complex social interactions that necessitate extensive behavioral coding systems (e.g., Dadds et al., 1994; O'Leary, Romanczyk, Kass, Dietz, & Santogrossi, 1971; Patterson, Ray, Shaw, & Cobb, 1969; Wahler, House, & Stambaugh, 1976).

The utility of behavioral observations in naturalistic and simulated settings is well illustrated in Ayllon, Smith, and Rogers' (1970) behavioral assessment of a young school-phobic girl. In this case study, impartial observers in the child's home monitored the stream of events occurring on school days in order to identify the actual school-phobic

behaviors and to determine the antecedent and consequent events associated with them. In this single-parent family, it was noted that the mother routinely left for work about one hour after the targeted girl (Valerie) and her siblings were to leave for school. Although the siblings left for school without incident, Valerie was observed clinging to her mother and refusing to leave the house and go to school. As described by Ayllon et al. (1970), "Valerie typically followed her mother around the house, from room to room, spending approximately 80 percent of her time within 10 feet of her mother. During these times there was little or no conversation" (p. 128). Given her refusal to go to school, the mother took Valerie to a neighbor's apartment for the day. However, when the mother attempted to leave for work, Valerie frequently followed her at a 10-foot distance. As a result, the mother had to return to the neighbor's apartment with Valerie in hand. This daily pattern was observed to end with the mother "literally running to get out of sight of Valerie" so she would not follow her to work. During the remainder of the day, it was observed that Valerie was allowed to do whatever she pleased: "Her day was one which would be considered ideal by many grade-school children—she could be outdoors and play as she chose all day long. No demands of any type were placed on her" (p. 129). Based on these observations, it appeared that Valerie's separation anxiety and refusal to attend school were related to her mother's attention and to the reinforcing environment of the neighbor's apartment where she could play all day.

However, because Valerie was also reported to be afraid of school itself, Ayllon et al. (1970) designed a simulated school setting in the home to determine the extent of anxiety or fear toward specific school-related tasks. (Obviously, observation in the school itself would have been desirable but was impossible because she refused to attend school.) Unexpected, little or no fear was evinced in the simulated setting; in fact, Valerie performed well and appeared to enjoy the school-related setting and homework tasks. In this case, these detailed behavioral observations were useful in ruling upon differential hypotheses related to school refusal. They led directly to a specific and efficacious treatment program based on shaping and differential reinforcement principles. The utility of behavioral observations for accurate assessment and treatment programming has been noted in numerous other case studies as well (e.g., Ollendick, 1995; Ollendick & Gruen, 1972; Smith & Sharpe, 1970).

A major disadvantage of behavioral observations in the natural environment is that the

target behavior may not occur during the designated observation periods. In such instances, simulated settings that occasion the target behaviors can be used. Simulated observations are especially helpful when the target behavior is of low frequency, when the target behavior is not observed in the naturalistic setting due to reactivity effects associated with being observed, or when the target behavior is difficult to observe in the natural environment due to practical constraints. Ayllon et al.'s (1970) use of a simulated school setting illustrated this approach under the latter conditions. A study by Matson and Ollendick (1976) illustrates this approach for low-frequency behaviors. In this study, parents reported that their children bit either the parent or siblings when they "were unable to get their way or were frustrated." Direct behavioral observations in the home confirmed the parental report, but it was necessary to observe the children for several hours prior to observing an occurrence of the behavior. Further, parents reported that their children were being "good" while the observers were present and that frequency of biting was much lower than its usual, "normal" rate. Accordingly, parents were trained in observation procedures and instructed to engage their children in play for four structured play sessions per day. During these sessions, parents were instructed to prompt biting behavior by deliberately removing a preferred toy. As expected, removal of favored toys in the structured situations resulted in increases in target behaviors, which were then eliminated through behavioral procedures. The structured, simulated play settings maximized the probability that biting would occur and that it could be observed and treated under controlled conditions.

It is often intimated that behavioral observation systems may not be suitable for more complex behavior problems, such as parent-child interactions. Sophisticated systems developed by Dumas (1989) and Dadds et al. (1994) to capture family interactions and processes suggest otherwise. For example, Dadds et al. (1994) developed the Family Anxiety Coding Schedule in order to measure anxious behavior in both child and parent, and the antecedents and consequences each provided the other to occasion anxiety in the other. This schedule was developed following the observation that children learned to process information about threat cues through interactions with their parents. More specifically, they observed that anxious children tended to view "neutral" situations as more threatening after discussing the situations with their parents than they did in the absence of such interactions. To learn more

about how this happened, Dadds and colleagues observed the moment-to-moment process whereby parents of anxious children influenced their children to change from a nonthreatened stance to an avoidant, threatened stance. To examine the interdependency of the parents and the child, they coded each family member's utterances in real time sequence so that conditional probabilities could be computed between different family members' behaviors. Using this system, they were able to show the process by which, and through which, the anxiety response was activated and maintained in the child. Thus a very complicated process of parent-child interactions was broken down into its constituent parts, recorded with a sophisticated observation system, and analyzed sequentially over time. Moreover, the observations suggested that, in this sample of overanxious children, anxiety did not exist solely "in" the child; rather, it existed in a context that was highly dependent upon parental influences. Such a demonstration illustrates the importance of contextual influences in understanding, assessing, and treating diverse child behavior disorders.

In sum, direct behavioral observation—either in the natural or simulated environment—provides valuable information for child behavioral assessment. When combined with information gathered through behavioral interviews, self- and other-reports, and self-monitoring, a comprehensive picture of children and their behaviors, as well as their controlling variables, is obtained. As with other assessment procedures, however, direct behavioral observation alone is not sufficient to meet the various behavioral assessment functions required for a thorough analysis of a child's problem behavior.

4.06.5 RESEARCH FINDINGS

As noted earlier, use of assessment instruments and procedures that have been empirically validated is one of the primary characteristics of child behavioral assessment. However, the role of conventional psychometric standards in evaluating child behavioral assessment procedures is a controversial one (e.g., Barrios & Hartman, 1986; Cone, 1981, 1986; Cone & Hawkins, 1977; Mash & Terdal, 1981). Given the theoretical underpinnings of child behavioral assessment and the basic assumptions regarding situational specificity and temporal instability of behavior, traditional psychometric standards would appear to be of little or no value. After all, how can behaviors thought to be under the control of specific

antecedent and consequent events be expected to be similar in different settings and at different times? Yet, if there is no consistency in behavior across settings and time, prediction of behavior is impossible and the generalizability of findings obtained from any one method of assessment would be meaningless. Such an extreme ideographic stance precludes meaningful assessment, except of highly discrete behaviors in very specific settings and at very specific points in time (Ollendick & Hersen, 1984).

Research findings suggest that it is not necessary totally to dismiss notions of cross-situational and cross-temporal consistency of behavior (e.g., Bem & Allen, 1974). Although a high degree of behavioral consistency cannot be expected, a moderate degree of behavioral consistency can be expected across situations that involve similar stimulus and response characteristics and are temporally related. When multimethod assessment procedures are used under these conditions, a modest relationship among the measures and a fair degree of predictability and generalizability can be expected. Under such circumstances, application of conventional psychometric standards to evaluation of child behavioral assessment procedures is less problematic and potentially useful (Cone, 1977; Ollendick & Hersen, 1984, 1993). The value of psychometric principles has already been demonstrated for certain classes of behavior when obtained through methods such as behavioral observation (e.g., Olweus, 1979), self-report (e.g., Ollendick, 1981), and other-report ratings (e.g., Cowen, Pederson, Barbikian, Izzo, & Trost, 1973). Further, when multiple methods of behavioral assessment have been used in the same studies, a modest degree of concurrent and predictive validity has been reported (e.g., Gresham, 1982).

It is beyond the scope of the present chapter to review specific research findings related to the reliability, validity, and clinical utility of the various procedures espoused in the multimethod approach. Nonetheless, brief mention will be made of specific directions of research and ways of enhancing the psychometric qualities of each procedure.

4.06.5.1 Behavioral Interviews

As noted by Evans and Nelson (1977), data based on retrospective reports obtained during the interview may possess both low reliability (agreement among individuals interviewed may differ and responses may vary over time) and low validity (reported information may not correspond to the "facts"). Such inaccurate or distorted recollections may result not only in

delayed clarification of the presenting complaints, but also in faulty hypotheses about causal agents and maintaining factors. For example, Chess, Thomas, and Birch (1966) reported that parents inaccurately reported certain behavior problems developed at times predicted by popular psychological theories. For example, problems with siblings were recalled to have begun with the birth of a younger sibling, and problems with dependency were reported to have begun when the mother became employed outside the home. In actuality, these behaviors were present prior to these events; nonetheless, they were "conveniently" recalled to have begun coincident with commonly accepted "life" points. In a similar vein, Schopler (1974) noted that many parents of autistic children inaccurately blame themselves for their child's problematic behaviors and that many therapists inadvertently "buy into" this notion that parents are to blame. Such scapegoating accomplishes little in the understanding, assessment, and treatment of the child's problematic behavior (Ollendick & Cerny, 1981).

While the reliability and validity of general information about parenting attitudes and practices are suspect, findings suggest parents and children can be reliable and valid reporters of current, specific information about problematic behaviors (e.g., Graham & Rutter, 1968; Gross, 1984; Herjanic, Herjanic, Brown, & Wheatt, 1973). The reliability and validity of the information are directly related to recency of behaviors being discussed and specificity of information requested. Thus, careful specification of precise behaviors and conditions under which they are occurring is more reliable and valid than vague descriptions of current behaviors or general recollections of early childhood events (Ciminero & Drabman, 1977). When the interview is conducted along such guidelines, it is useful in specifying behaviors of clinical interest and in determining appropriate therapeutic interventions. As we have noted, however, it is only the first step in the ongoing, hypothesis-generating process that is characteristic of child behavioral assessment.

4.06.5.2 Ratings and Checklists

As with behavioral interviews, issues related to reliability and validity are also relevant to ratings and checklists. Cronbach (1960) has noted that the psychometric quality of rating scales is directly related to the number and specificity of the items rated. Further, O'Leary and Johnson (1986) have identified four factors associated with item-response characteristics

and raters that enhance reliability and validity of such scales: (i) the necessity of using clearly defined reference points on the scale (i.e., estimates of frequency, duration, or intensity), (ii) the inclusion of more than two reference points on the scale (i.e., reference points that quantify the behavior being rated), (iii) a rater who has had extensive opportunities for observing the child being rated, and (iv) more than one rater who has equal familiarity with the child.

The rating forms and checklists described earlier (e.g., Revised Behavior Problem Checklist, Child Behavior Checklist, Behavior Assessment System for Children, the Louisville Fear Survey Schedule for Children, and the Home Situations Questionnaire) incorporate these item and response characteristics and are generally accepted as reliable and valid instruments. For example, the interrater reliability of the Revised Behavior Problem Checklist is quite high when raters are equally familiar with the children being rated and when ratings are provided by raters within the same setting (Quay, 1977; Quay & Peterson, 1983). Further, stability of these ratings has been reported over two-week and one-year intervals. These findings have been reported for teachers in the school setting and parents in the home setting. However, when ratings of teachers are compared to those of parents, interrater reliabilities are considerably lower. While teachers seem to agree with other teachers, and one parent tends to agree with the other parent, there is less agreement between parents and teachers. Such differences may be due to differential perceptions of behavior by parents and teachers or to the situational specificity of behavior, as discussed earlier (also see Achenbach et al., 1987). These findings support the desirability of obtaining information about the child from as many informants and from as many settings as possible.

The validity of the Revised Behavior Problem Checklist has also been demonstrated in numerous ways. It has been shown to distinguish clinic-referred children from nonreferred children, and to be related to psychiatric diagnosis, other measures of behavioral deviance, prognosis, and differential effectiveness of specific treatment strategies (see Ollendick & Cerny, 1981, for a discussion of these findings).

Findings similar to these have been reported for the Child Behavior Checklist, Behavior Assessment System for Children, Louisville Fear Survey Schedule, and the Home Situations Questionnaire. These rating forms and checklists, as well as others, have been shown to possess sound psychometric qualities and to be clinically useful. They not only provide mean-

ingful data about the child's adaptive and problem behaviors but are also useful in orienting parents, teachers, and significant others to specific problem or asset areas and in alerting them to observe and record specific behaviors accurately and validly.

4.06.5.3 Self-report Instruments

Of the various methods used in child behavioral assessment, the self-report method has received the least empirical support, although this picture is rapidly changing. As noted earlier, child behavioral assessors initially eschewed use of self-report instruments, largely on the basis of their suspected low reliability and validity. As we have noted, however, data from self-report instruments can be meaningfully used to understand and describe the child, plan treatment, and evaluate treatment outcome.

As with interview and checklist or rating data, self-report of specific behaviors (including cognitions and affects) and events is more reliable and valid than more general, global reports of life experiences. Such self-reports of specific states can be used to identify discrete components of more general constructs (e.g., determining the exact fears of a phobic child and the exact situations that are associated with withdrawn behavior in an unassertive child). Illustratively, Scherer and Nakamura's (1968) Fear Survey Schedule for Children and its revision. (Ollendick, 1983b) can be used to pinpoint specific fears and classes of fear. Further, this instrument has been shown to be reliable over time, to possess high internal consistency and a meaningful and replicable factor structure, to distinguish between phobic and nonphobic children, and to discriminate among subtypes of phobic youngsters within a particular phobic group (Ollendick & Mayer, 1984; Ollendick, King, & Yule, 1994).

Clearly, more research is needed in this area before routine use of self-report instruments can be endorsed. Nonetheless, instruments that measure specific aspects of behavior such as anxiety or depression rather than global traits hold considerable promise for child behavioral assessment.

4.06.5.4 Self-monitoring

In self-monitoring, children observe their own behavior and then systematically records its occurrence. As with other measures, concerns related to the reliability and validity of this method exist. What is the extent of interobserver agreement between children who are instructed to monitor their own behavior

and objective observers? How accurate are children in recording occurrences of behavior? How reactive is the process of self-monitoring?

The literature in this area is voluminous. Even though all necessary studies have not been conducted, the findings are in general agreement. First, children as young as seven years of age can be trained to be reliable and accurate recorders of their own behavior. However, the specific behaviors should be clearly defined, prompts to self-record should be available, and reinforcement for self-monitoring should be provided. Under such conditions, children's recordings closely approximate those obtained from observing adults. For example, in a study examining the effects of self-monitoring and self-administered overcorrection in the treatment of nervous tics in children, Ollendick (1981) showed that 8–10-year-old children who were provided clear prompts to self-record highly discrete behaviors were able to do so reliably. Estimates of occurrence closely paralleled those reported by parents and teachers, even though children were unaware that these adults were recording their nervous tics. In another study, Ackerman and Shapiro (1985) demonstrated the accuracy of self-monitoring by comparing self-recorded data with a permanent product measure (the number of units produced in a work setting). Again, accuracy of self-monitoring was confirmed.

Second, self-monitoring may result in behavior change due to the self-observation process and result in altered estimates of target behaviors. This effect is known as reactivity. Numerous factors have been shown to influence the occurrence of reactivity: specific instructions, motivation, goal-setting, nature of the self-recording device, and the valence of the target behavior (e.g., Nelson, 1977, 1981). Among the more important findings are that desirable behaviors (e.g., study habits, social skills) may increase while undesirable behaviors (e.g., nervous tics, hitting) tend to decrease following self-monitoring, and that the more obtrusive the self-recording device, the greater the behavior change. For example, Nelson, Lipinski, and Boykin (1978) found that hand-held counters produced greater reactivity than belt-worn counters. Holding a counter in one's hand was viewed as more obtrusive, contributing to increased reactivity. Reactivity is a concern in the assessment process because it affects the actual occurrences of behavior. However, if one is aware of the variables that contribute to reactive effects, self-monitoring can be used as a simple and efficient method for data collection (Shapiro, 1984).

In short, self-monitoring has been found to be useful in the assessment of a wide range of child behavior problems across a wide variety of settings. When issues related to the reliability, accuracy, and reactivity of measurement are addressed, self-monitoring represents another clinically useful strategy that is highly efficient and clinically useful.

4.06.5.5 Behavioral Observation

As with other assessment strategies, behavioral observation procedures must possess adequate psychometric qualities and be empirically validated before their routine use can be endorsed. Although early behaviorists accepted the accuracy of behavioral observations based on their face validity, subsequent investigators enumerated a variety of problems associated with their reliability, validity, and clinical utility (e.g., Johnson & Bolstad, 1973; Kazdin, 1977). These problems include the complexity of the observation code, the exact recording procedures to be used (e.g., frequency counts, time sampling, etc.), observer bias, observer drift, and the reactive nature of the observation process itself (see Barton & Ascione, 1984, for further discussion of these issues). Our experience suggests that the greatest threat to the utility of observational data comes from the reactive nature of the observational process itself, especially when the observer is present in the natural setting. It is well known that the presence of an observer affects behavior, usually in socially desirable directions. We have found two strategies to be useful in reducing such reactive effects: recruiting and training observer-coders already present in the natural setting (e.g., a teacher or parent), and if this is not possible, planning extended observations so children can habituate to the observers and so that the effects of reactivity will diminish. However, in the latter instance, it should be noted that several sessions of observations may be required, since reactive effects have been observed to be prolonged (Johnson & Lobitz, 1974). Reactive effects, combined with the aforementioned practical issues of personnel, time, and resources, have led us to place greater emphasis on recruiting observer-coders already in the children's natural environment or training children themselves to record their own behavior.

In brief, behavioral observations are the most direct and least inferential method of assessment. Even though a variety of problems related to their reliability and validity have been commented upon, behavioral observations are highly useful strategies and represent the hall-

mark of child behavioral assessment. Whenever possible, behavioral observations in the natural setting should be obtained.

4.06.6 FUTURE DIRECTIONS

A number of directions for future research and development in child behavioral assessment may be evident to the reader. What follows is our attempt to highlight those areas that appear most promising and in need of greater articulation.

4.06.6.1 Developmental Factors

First, it seems to us that greater attention must be given to developmental factors as they affect the selection of child behavioral assessment procedures. Although we have argued that these procedures should be developmentally sensitive, child behavioral assessors have frequently not attended to, or have ignored, this recommendation. As we noted earlier, the most distinguishing characteristic of children is developmental change. Such change encompasses basic biological growth and maturity as well as affective, behavioral, and cognitive fluctuations that may characterize children at different age levels. While the importance of accounting for developmental level when assessing behavior may be obvious, ways of integrating developmental concepts and principles into child behavioral assessment are less evident. Edelbrock (1984) has noted three areas for the synthesis of developmental and behavioral principles: (i) use of developmental fluctuations in behavior to establish normative baselines of behavior, (ii) determination of age and gender differences in the expression and covariation of behavioral patterns, and (iii) study of stability and change in behavior over time. Clearly, these areas of synthesis and integration are in their infancy and in need of greater articulation (e.g., Harris & Ferrari, 1983; Ollendick & Hersen, 1983; Rutter & Garmezy, 1983; Sroufe & Rutter, 1984).

Recently, Ollendick and King (1991) addressed this developmental-behavioral synthesis in some detail. In reference to normative data, they suggested that such information could be used to determine which behavior problems represent clinically significant areas of concern, examine appropriateness of referral, and evaluate efficacy of interventions. Essentially, this normative-developmental perspective emphasizes the central importance of change over time and the need for relevant norms against which children can be compared,

both at the time of assessment and following treatment. Another way in which developmental principles can be integrated into ongoing child behavioral assessment is to identify age differences in the relations or patterns among behaviors (Edelbrock, 1984). Ollendick and King (1991) have shown such patterning of behavior across development for a number of measures, including diagnostic interviews, self- and other-report instruments, and behavioral coding systems.

Finally, developmental principles can be useful in child behavioral assessment in our attempts to examine and understand continuity and discontinuity of certain behavioral patterns. Basically, this issue can be addressed from two vantage points, a descriptive one and an explanatory one. From a descriptive standpoint, we are interested in determining whether a behavior or set of behaviors seen at one point in time can be described in the same way at another point in time. If it can be described in the same way, descriptive continuity is said to exist; if it cannot, descriptive discontinuity is said to obtain (Lerner, 1986). We are simply asking, does the behavior look the same or different? Does it take the same form over time? For example, if 4-year-old, 8-year-old, and 12-year-old children all emitted the same behaviors to gain entry in a social group, we would conclude that descriptive continuity exists for social entry behavior.

For the most part, it has been shown that the expression and patterning of a large number of behaviors change across development and that descriptive discontinuity is more likely the case (Ollendick & King, 1991). Changes in behavior observed with development can, of course, occur for many different reasons. If the same explanations are used to account for behavior over time, then that behavior is viewed as involving unchanging laws or rules and explanatory continuity is said to exist. However, if different explanations are used to account for changes in behavior over time, explanatory discontinuity prevails (Lerner, 1986). For the most part, behaviorally oriented theorists and clinicians maintain changes over time are due to a set of learning principles that are largely the same across the child's life span. No new principles or laws are needed as the child grows. Developmental theorists, on the other hand, maintain a progressive differentiation of the organism which suggests a different set of principles be invoked across different stages of development. Unfortunately, the evidence on explanatory continuity versus discontinuity is scarce; the jury is out on these issues. Much work remains to be done in this area; however, as Ollendick and King (1991) note, the

emergence of “developmental-behavioral assessment” is on the horizon.

4.06.6.2 The Utility of the Multimethod Approach at Different Age Levels

Second, and somewhat related to the first area, greater attention must be focused on the incremental validity of the multimethod approach when used for children of varying ages. Throughout this chapter, we have espoused a multimethod approach consisting of interviews, self- and other-reports, self-monitoring, and behavioral observations. Some of these procedures may be more appropriate at some age levels than others. Further, the psychometric properties of these procedures may vary with age. For example, self-monitoring requires the ability to compare one's own behavior against a standard and accurately to judge occurrence or nonoccurrence of targeted events and behaviors. Most children below six years of age lack the requisite ability to self-monitor and may not profit from such procedures. In fact, the limited research available suggests self-monitoring may be counter-productive when used with young children, resulting in confusion and impaired performance (e.g., Higa, Thorp, & Calkins, 1978). These findings suggest that self-monitoring procedures are better suited for children who possess sufficient cognitive abilities to benefit from their use (Shapiro, 1984). In a similar vein, age-related variables place constraints on use of certain self-report and sociometric measures with young children. It has often been noted that sociometric devices must be simplified and presented in pictorial form to children under six years of age (Hops & Lewin, 1984). The picture-form sociometric device provides young children with a set of visual cues regarding children to be rated and, of course, does not require them to read names of children being rated. The roster-and-rating method, used so frequently with older children, is simply not appropriate with younger children. Ollendick and Hersen (1993) review additional age-related findings for other procedures and suggest caution in using these procedures without due regard for their developmental appropriateness and related psychometric properties.

If certain procedures are found to be less reliable or valid at different age levels, their indiscriminate use with children can not be endorsed. Inasmuch as these strategies are found to be inadequate, the combination of them in a multimethod approach would serve only to compound their inherent limitations

(Mash & Terdal, 1981). The *sine qua non* of child behavioral assessment is that the procedures be empirically validated. In addition, the different procedures might vary in terms of their treatment utility across different ages. Treatment utility refers to the degree to which assessment strategies are shown to contribute to beneficial treatment outcomes (Hayes et al., 1987). More specifically, treatment utility addresses issues related to the selection of specific target behaviors and to the choice of specific assessment strategies. For example, we might wish to examine the treatment utility of using self-report questionnaires to guide treatment planning, above and beyond that provided by direct behavioral observation of children who are phobic of social encounters. All children could complete a fear schedule and be observed in a social situation, but the self-report data for only half of the children would be made available for treatment planning. If the children for whom self-reports were made available improved more than those whose treatment plans were based solely on behavioral observations, then the treatment utility of using self-report data would be established (for this problem with this age child). In a similar fashion, the treatment utility of interviews, role plays, and other devices could be evaluated (Hayes et al., 1987). Of course, it would be important to examine treatment utility from a developmental perspective as well. Certain procedures might be shown to possess incremental validity at one age but not another. Although the concept of treatment utility is relatively new, it shows considerable promise as a strategy to evaluate the incremental validity of our multimethod assessment approach. We should not necessarily assume that “more” assessment is “better” assessment.

4.06.6.3 Cultural Sensitivity

Considerable energy must be directed to the development of child behavioral assessment methods that are culturally sensitive. Numerous observers have called attention to the internationalization of the world and the “browning of America” (e.g., Malgady, Rogler, & Constantino, 1987; Vasquez Nuttall, DeLeon, & Del Valle, 1990). In reference to this chapter, these developments suggest that the assessment process is increasingly being applied to non-Caucasian children for whom English is not the primary language in America, and that many procedures developed in America and other Western countries are being applied, sometimes indiscriminately, in other countries as well. Development of assessment procedures that are

culture-fair (and language-fair) is of utmost importance. Of course, many cultural issues need to be considered in the assessment process. Cultural differences may be expressed in child-rearing practices, family values, parental expectations, communication styles, nonverbal communication patterns, and family structure and dynamics (Vasquez et al., 1990). As an example, behaviors characteristic of ethnic minority children may be seen as emotionally or behaviorally maladaptive by persons who have little or no appreciation for cultural norms (e.g., Prewitt-Diaz, 1989). Thus, cultural differences (biases?) are likely to occur early in the assessment process. Fortunately, Vasquez-Nuttall, Sanchez, Borrás Osorio, Nuttall, & Varvogil (1996) have suggested several steps that can be taken to minimize cultural biases in the assessment process.

Vasquez et al. (1996) have offered the following suggestions: (i) include extended family members in the information-gathering process; (ii) use interpreters, if necessary, in interviewing the child and family members; (iii) familiarize oneself with the culture of specific groups; and (iv) use instruments that have been translated into the native language of the children and for which norms are available for specific ethnic groups. With regard to this latter recommendation, significantly greater progress has been witnessed for the translation component than the establishment of well-standardized normative information. For example, while the Conners' Parent Rating Scales (Conners, 1985) and Conners' Teacher Rating Scale (Conners, 1985) have been translated into Spanish and other languages, group norms are lacking and the reliability and validity of the translations have not been examined systematically. Similarly, the Fear Survey Schedule for Children-Revised (Ollendick, 1983b) has been translated into over 10 languages, yet normative data are lacking and the psychometric properties of the instrument have not been fully explored or established. In sum, a clear challenge before us in the years ahead is to attend to important cultural factors that impinge on our assessment armamentarium, and to develop and promulgate culturally sensitive methods that are developmentally appropriate and empirically validated.

4.06.6.4 Measures of Cognitive and Affective Processes

More effort must be directed toward the development of culturally relevant, developmentally sensitive, and empirically validated procedures for assessment of cognitive and affective processes in children. In recent years,

child behavioral assessors have become increasingly interested in the relation of children's cognitive and affective processes to observed behaviors. The need for assessment in this area is further evidenced by the continued increase of cognitive-behavioral treatment procedures with children, a trend first observed in the late 1970s and early 1980s (e.g., Kendall, Pellegrini, & Urbain, 1981; Meador & Ollendick, 1984). As noted by Kendall et al. (1981), there is a particularly pressing need to develop procedures that can examine the very cognitions and processes that are targeted for change in these intervention efforts. For example, the reliable and valid assessment of self-statements made by children in specific situations would facilitate the empirical evaluation of cognitive-behavioral procedures such as self-instructional training and cognitive restructuring (cf. Zatz & Chassin, 1983; Stefanek, Ollendick, Baldock, Francis, & Yaeger, 1987).

4.06.6.5 The Role of the Child

We must concentrate additional effort on the role of the child in child behavioral assessment. All too frequently, "tests are administered to children, ratings are obtained on children, and behaviors are observed in children" (Ollendick & Hersen, 1984, p. ix). This process views the child as a passive responder, someone who is largely incapable of actively shaping and determining behaviors of clinical relevance. Although examination of these organismic variables is only beginning, it would appear that concerted and systematic effort must be directed to their description and articulation. For example, children's conceptions of their own behavior is a critical area of further study. To what causes do children attribute aggressive or withdrawn behavior in themselves or in their peers? Are there aggregated trends in these attributions? Do they differ by culture? Do causal attributions (as well as self-efficacy and outcome expectancies) mediate treatment outcomes? Again, are there age-related effects for these effects or culturally relevant effects? The answers to these questions are of both theoretical interest and applied clinical significance.

The process described above also implies that child behavior (problematic or otherwise) occurs in a vacuum, and that the perceptions and behaviors of referral sources (parents, teachers) and characteristics of the environments in which behavior occurs are somehow less critical to assess. Recent efforts to develop reliable methods for assessing parent-child interactions are indicative of an increased awareness of the need to broaden the scope of assessment to include specific individuals with

whom, and environments in which, child behavior problems commonly occur (cf. Dadds et al., 1994; Dumas, 1989; Greene, 1995, 1996; Ollendick, 1996). However, much additional work remains to be done in this area.

4.06.6.6 Ethical Guidelines

Finally, we must continue to focus our attention on ethical issues in child behavioral assessment. A number of ethical issues regarding children's rights, proper and legal consent, professional judgment, and social values are raised in the routine practice of child behavioral assessment (Rekers, 1984). Are children capable of granting full and proper consent to a behavioral assessment procedure? At what age and in what cultures are children competent to give such consent? Is informed consent necessary? Or might not informed consent be impossible, impractical, or countertherapeutic in some situations? What ethical guidelines surround the assessment procedures to be used? Current professional guidelines suggest our procedures should be reliable, valid, and clinically useful. Do the procedures suggested in this chapter meet these professional guidelines? What are the rights of parents and of society? It should be evident from these questions that a variety of ethical issues persists. Striking a balance between the rights of parents, society, and children is no easy matter but is one that takes on added importance in the increasingly litigious society of the USA.

In short, future directions of child behavioral assessment are numerous and varied. Even though a technology for child behavioral assessment has evolved and is in force, we need to begin to explore the issues raised before we can conclude the procedures are maximally productive and in the best interests of children throughout the world.

4.06.7 SUMMARY

Child behavioral assessment strategies have been slow to evolve. Only recently has the chasm between child behavior therapy and child behavioral assessment been narrowed. Increased awareness of the importance of developing assessment procedures that provide an adequate representation of child behavior disorders has spurred research into assessment procedures and spawned a plethora of child behavioral assessment techniques. The growing sophistication of child behavior assessment is witnessed by the appearance of self- and other-report strategies that are beginning to take into account developmental, social, and cultural

influences as well as cognitive and affective mediators of overt behavior. At the same time, attention to psychometric properties of assessment procedures has continued.

Certain theoretical assumptions guide child behavioral assessment. Foremost among these is the premise that behavior is a function of situational determinants and not a sign of underlying personality traits. To assess adequately the situational determinants and to obtain as complete a picture of the child as is possible, a multimethod assessment approach is recommended, utilizing both direct and indirect methods of assessment. Direct methods include self-monitoring as well as behavioral observation by trained observers in naturalistic or simulated analogue settings. Indirect measures include behavioral interviewing and self- and other-report measures. These sources of information are considered indirect ones because they involve retrospective reports of previous behavior.

Even though direct behavioral observation remains the hallmark of child behavioral assessment, information from these other sources is considered not only valuable but integral in the understanding and subsequent treatment of child behavior disorders. Hence, whereas identification and specification of discrete target behaviors were once considered sufficient, current child behavioral assessment involves serious consideration and systematic assessment of cognitive and affective aspects of the child's behavior and of developmental, social, and cultural factors that influence the child, as well as direct observation of the problematic behavior *in situ*.

Several areas of future research remain. These include clearer specification of developmental variables, a closer examination of the utility of the multimethod approach at different age levels, the influence of culture and the need for models of assessment that take cultural forces into consideration, development of specific measures to examine cognitive and affective processes in children, articulation of the role of the child in child behavioral assessment, and continued development of ethical guidelines. While the basis for a technology of child behavioral assessment exists, considerable fine-tuning remains to be done. Child behavioral assessment is at a critical crossroad in its own development; which path it takes will determine its long-term future.

4.06.8 REFERENCES

- Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs*, 80, 1-37.

- Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T. M. (1991b). *Manual for the Teacher Report Form and 1991 Profile*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T. M. (1991c). *Manual for the Youth Self-Report and 1991 Profile*. Burlington, VT: University of Vermont Department of Psychiatry.
- Achenbach, T. M., & Edelbrock, C. S. (1989). Diagnostic, taxonomic, and assessment issues. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed., pp. 53–69). New York: Plenum.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101, 213–232.
- Ackerman, A. M., & Shapiro, E. S. (1985). Self-monitoring and work productivity with mentally retarded adults. *Journal of Applied Behavior Analysis*, 17, 403–407.
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Ayllon, T., Smith, D., & Rogers, M. (1970). Behavioral management of school phobia. *Journal of Behavior Therapy and Experimental Psychiatry*, 1, 125–138.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191–215.
- Barkley, R. A. (1981). *Hyperactive children: A handbook for diagnosis and treatment*. New York: Guilford Press.
- Barkley, R. A. (1987). *Defiant children: A clinician's manual for parent training*. New York: Guilford Press.
- Barkley, R. A., & Edelbrock, C. S. (1987). Assessing situational variation in children's behavior problems: The home and school situations questionnaires. In R. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 3, pp. 157–176). Greenwich, CT: JAI Press.
- Barkley, R. A., Karlsson, I., Strzelecki, E., & Murphy, J. (1984). Effects of age and Ritalin dosage on the mother-child interactions of hyperactive children. *Journal of Consulting and Clinical Psychology*, 52, 750–758.
- Barrios, B., & Hartmann, D. P. (1986). The contributions of traditional assessment: Concepts, issues, and methodologies. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 81–110). New York: Guilford Press.
- Barton, E. J., & Ascione, F. R. (1984). Direct observations. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 166–194). New York: Pergamon.
- Bem, D. I., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, 81, 506–520.
- Bornstein, P. H., Bornstein, M. T., & Dawson, B. (1984). Integrated assessment and treatment. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 223–243). New York: Pergamon.
- Campbell, S. B. (1989). Developmental perspectives in child psychopathology. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed., pp. 5–28). New York: Plenum.
- Cantwell, D. P. (1983). Childhood depression: A review of current research. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 5, pp. 39–93). New York: Plenum.
- Chess, S., Thomas, A., & Birch, H. G. (1966). Distortions in developmental reporting made by parents of behaviorally disturbed children. *Journal of the American Academy of Child Psychiatry*, 5, 226–231.
- Ciminero, A. R., & Drabman, R. S. (1977). Current developments in the behavioral assessment of children. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 1, pp. 47–82). New York: Plenum.
- Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 8, 411–426.
- Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and taxonomy. *Behavior Therapy*, 9, 882–888.
- Cone, J. D. (1981). Psychometric considerations. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (2nd ed., pp. 38–68). Elmsford, NY: Pergamon.
- Cone, J. D. (1986). Ideographic, nomothetic, and related perspectives in behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 111–128). New York: Guilford Press.
- Cone, J. D., & Hawkins, R. P. (Eds.) (1977). *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel.
- Conners, C. K. (1985). *The Conners rating scales: Instruments for the assessment of childhood psychopathology*. Unpublished manuscript, Children's Hospital National Medical Center, Washington, DC.
- Cowen, E. L., Pederson, A., Barbighian, H., Izzo, L. D., & Trost, M. A. (1973). Long-term follow-up of early-detected vulnerable children. *Journal of Consulting and Clinical Psychology*, 41, 438–445.
- Cronbach, L. J. (1960). *Essentials of psychological testing*. New York: Harper & Row.
- Dadds, M. R., Rapee, R. M., & Barrett, P. M. (1994). Behavioral observation. In T. H. Ollendick, N. J. King, & W. Yule (Eds.), *International handbook of phobic and anxiety disorders in children and adolescents* (pp. 349–364). New York: Plenum.
- Deluty, R. H. (1979). Children's Action Tendency Scale: A self-report measure of aggressiveness, assertiveness, and submissiveness in children. *Journal of Consulting and Clinical Psychology*, 41, 1061–1071.
- Dong, Q., Yang, B., & Ollendick, T. H. (1994). Fears in Chinese children and adolescents and their relations to anxiety and depression. *Journal of Child Psychology and Psychiatry*, 35, 351–363.
- Doyle, A., Ostrander, R., Skare, S., Crosby, R. D., & August, G. J. (1997). Convergent and criterion-related validity of the Behavior Assessment System for Children-Parent Rating Scale. *Journal of Clinical Child Psychology*, 26, 276–284.
- Dumas, J. E. (1989). Interact: A computer-based coding and data management system to assess family interactions. In R. J. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 3, pp. 177–202). Greenwich, CT: JAI Press.
- Edelbrock, C. S. (1984). Developmental considerations. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 20–37). Elmsford, NY: Pergamon.
- Evans, I. M., & Nelson, R. O. (1977). Assessment of child behavior problems. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 603–681). New York: Wiley-Interscience.
- Finch, A. J., Nelson, W. M., III, & Moss, J. H. (1983). Stress inoculation for anger control in aggressive children. In A. J. Finch, W. M. Nelson, & E. S. Ott (Eds.), *Cognitive-behavioral procedures with children: A practical guide* (pp. 148–205). Newton, MA: Allyn & Bacon.
- Finch, A. J., & Rogers, T. R. (1984). Self-report instruments. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 106–123). Elmsford, NY: Pergamon.
- Goldfried, M. R., & Kent, R. N. (1972). Traditional versus

- behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, 77, 409–420.
- Graham, P., & Rutter, M. (1968). The reliability and validity of the psychiatric assessment of the child—II. Interview with the parents. *British Journal of Psychiatry*, 114, 581–592.
- Greene, R. W. (1995). Students with ADHD in school classrooms: Teacher factors related to compatibility, assessment, and intervention. *School Psychology Review*, 24, 81–93.
- Greene, R. W. (1996). Students with ADHD and their teachers: Implications of a goodness-of-fit perspective. In T. H. Ollendick & R. J. Prinz (Eds.), *Advances in clinical child psychology* (Vol. 18, pp. 205–230). New York: Plenum.
- Greene, R. W., & Ollendick, T. H. (in press). Behavioral assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (3rd ed.). Boston: Allyn & Bacon.
- Gresham, F. M. (1982). Social interactions as predictors of children's likability and friendship patterns: A multiple regression analysis. *Journal of Behavioral Assessment*, 4, 39–54.
- Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system manual*. Circle Pines, MN: American Guidance Service.
- Gross, A. M. (1984). Behavioral interviewing. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 61–79). Elmsford, NY: Pergamon.
- Harris, S. L., & Ferrari, M. (1983). Developmental factors in child behavior therapy. *Behavior Therapy*, 14, 54–72.
- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1986). Evaluating the quality of behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 463–503). New York: Guilford.
- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist*, 42, 963–974.
- Herjanic, B., Herjanic, M., Brown, F., & Wheatt, T. (1973). Are children reliable reporters? *Journal of Abnormal Child Psychology*, 3, 41–48.
- Higa, W. R., Tharp, R. G., & Calkins, R. P. (1978). Developmental verbal control of behavior: Implications for self-instructional testing. *Journal of Experimental Child Psychology*, 26, 489–497.
- Holmes, F. B. (1936). An experimental investigation of a method of overcoming children's fears. *Child Development*, 1, 6–30.
- Hops, H., & Lewin, L. (1984). Peer sociometric forms. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 124–147). New York: Pergamon.
- Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observations: Some problems and solutions for field research. In L. A. Hammerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice* (pp. 7–67). Champaign, IL: Research Press.
- Johnson, S. M., & Lobitz, G. K. (1974). Parental manipulation of child behavior in home observations. *Journal of Applied Behavior Analysis*, 1, 23–31.
- Jones, M. C. (1924). The elimination of children's fears. *Journal of Experimental Psychology*, 7, 382–390.
- Jones, R. R., Reid, J. B., & Patterson, G. R. (1975). Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 3, pp. 42–95). San Francisco: Jossey-Bass.
- Kanfer, F. H., & Phillips, J. S. (1970). *Learning foundations of behavior therapy*. New York: Wiley.
- Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, 4, 7–14.
- Kendall, P. C., & Hollon, S. D. (Eds.) (1980). *Cognitive-behavioral intervention: Assessment methods*. New York: Academic Press.
- Kendall, P. C., Pellegrini, D. S., & Urbain, E. S. (1981). Approaches to assessment for cognitive-behavioral interventions with children. In P. C. Kendall & S. D. Hollon (Eds.), *Assessment strategies for cognitive-behavioral interventions* (pp. 227–286). New York: Academic Press.
- Kovacs, M. (1985). Children's Depression Inventory (CDI). *Psychopharmacology Bulletin*, 21, 995–998.
- Kunzelman, H. D. (Ed.) (1970). *Precision teaching*. Seattle, WA: Special Child Publications.
- Lease, C. A., & Ollendick, T. H. (1993). Development and psychopathology. In A. S. Bellack & M. Hersen (Eds.), *Psychopathology in adulthood* (pp. 89–102). Boston: Allyn & Bacon.
- Lerner, R. M. (1986). *Concepts and theories of human development* (2nd ed.). New York: Random House.
- Linehan, M. (1977). Issues in behavioral interviewing. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology* (pp. 30–51). New York: Brunner/Mazel.
- Malgady, R., Rogler, L., & Constantino, G. (1987). Ethnocultural and linguistic bias in mental health evaluation of Hispanics. *American Psychologist*, 42, 228–234.
- Mash, E. J., & Terdal, L. G. (1981). Behavioral assessment of childhood disturbance. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders* (pp. 3–76). New York: Guilford Press.
- Mash, E. J., & Terdal, L. G. (Eds.) (1989). *Behavioral assessment of childhood disorders* (2nd ed.). New York: Guilford Press.
- Mash, E. J., & Terdal, L. G. (1989). Behavioral assessment of childhood disturbance. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders* (2nd ed., pp. 3–65). New York: Guilford Press.
- Matson, J. L., & Ollendick, T. H. (1976). Elimination of low-frequency biting. *Behavior Therapy*, 7, 410–412.
- McConaughy, S. H. (1996). The interview process. In M. J. Breen & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/behavioral disorders: A handbook for school-based practitioners* (pp. 181–224). Austin, TX: ProEd.
- McMahon, R. J. (1984). Behavioral checklists and rating forms. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 80–105). Elmsford, NY: Pergamon.
- Meador, A. E., & Ollendick, T. H. (1984). Cognitive behavior therapy with children: An evaluation of its efficacy and clinical utility. *Child and Family Behavior Therapy*, 6, 25–44.
- Meichenbaum, D. H. (1977). *Cognitive-behavior modification*. New York: Plenum.
- Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. (1972). Comparison of reciprocal inhibition, psychotherapy, and waiting list control for phobic children. *Journal of Abnormal Psychology*, 79, 269–279.
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, 80, 252–283.
- Nelson, R. O. (1977). Methodological issues in assessment via self-monitoring. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology* (pp. 217–240). New York: Brunner/Mazel.
- Nelson, R. O. (1981). Theoretical explanations for self-monitoring. *Behavior Modification*, 5, 3–14.

- Nelson, R. O., Lipinski, D. P., & Boykin, R. A. (1978). The effects of self-recorder training and the obtrusiveness of the self-monitoring device on the accuracy and reactivity of self-monitoring. *Behavior Therapy*, 9, 200-208.
- Nelson, W. M., III, & Finch, A. J., Jr. (1978). *The new children's inventory of anger*. Unpublished manuscript, Xavier University, OH.
- Novick, J., Rosenfeld, E., Bloch, D. A., & Dawson, D. (1966). Ascertaining deviant behavior in children. *Journal of Consulting and Clinical Psychology*, 30, 230-238.
- O'Leary, K. D., & Johnson, S. B. (1986). Assessment and assessment of change. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of children* (3rd ed., pp. 423-454). New York: Wiley.
- O'Leary, K. D., Romanczyk, R. G., Kass, R. E., Dietz, A., & Santogrossi, D. (1971). *Procedures for classroom observations of teachers and parents*. Unpublished manuscript, State University of New York at Stony Brook.
- Ollendick, T. H. (1981). Self-monitoring and self-administered overcorrection: The modification of nervous tics in children. *Behavior Modification*, 5, 75-84.
- Ollendick, T. H. (1983a). Development and validation of the Children's Assertiveness Inventory. *Child and Family Behavior Therapy*, 5, 1-15.
- Ollendick, T. H. (1983b). Reliability and validity of the Revised-Fear Survey Schedule for Children (FSSC-R). *Behaviour Research and Therapy*, 21, 685-692.
- Ollendick, T. H. (1995). Cognitive-behavioral treatment of panic disorder with agoraphobia in adolescents: A multiple baseline design analysis. *Behavior Therapy*, 26, 517-531.
- Ollendick, T. H. (1996). Violence in society: Where do we go from here? (Presidential address). *Behavior Therapy*, 27, 485-514.
- Ollendick, T. H., & Cerny, J. A. (1981). *Clinical behavior therapy with children*. New York: Plenum.
- Ollendick, T. H., & Greene, R. W. (1990). Behavioral assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed., pp. 403-422). Elmsford, NY: Pergamon.
- Ollendick, T. H., & Gruen, G. E. (1972). Treatment of a bodily injury phobia with implosive therapy. *Journal of Consulting and Clinical Psychology*, 38, 389-393.
- Ollendick, T. H., & Hersen, M. (Eds.) (1983). *Handbook of child psychopathology*. New York: Plenum.
- Ollendick, T. H., & Hersen, M. (Eds.) (1984). *Child behavioral assessment: Principles and procedures*. New York: Pergamon.
- Ollendick, T. H., & Hersen, M. (1993). Child and adolescent behavioral assessment. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child and adolescent behavioral assessment* (pp. 3-14). New York: Pergamon.
- Ollendick, T. H., & King, N. J. (1991). Developmental factors in child behavioral assessment. In P. R. Martin (Ed.), *Handbook of behavior therapy and psychological science: An integrative approach* (pp. 57-72). New York: Pergamon.
- Ollendick, T. H., & King, N. J. (1994). Assessment and treatment of internalizing problems: The role of longitudinal data. *Journal of Consulting and Clinical Psychology*, 62, 918-927.
- Ollendick, T. H., King, N. J., & Frary, R. B. (1989). Fears in children and adolescents in Australia and the United States. *Behaviour Research and Therapy*, 27, 19-26.
- Ollendick, T. H., King, N. J., & Yule, W. (Eds.) (1994). *International handbook of phobic and anxiety disorders in children*. Boston: Allyn & Bacon.
- Ollendick, T. H., Matson, J. L., & Hetsel, W. J. (1985). Fears in children and adolescents: Normative data. *Behaviour Research and Therapy*, 23, 465-467.
- Ollendick, T. H., & Mayer, J. (1984). School phobia. In S. M. Turner (Ed.), *Behavioral treatment of anxiety disorders* (pp. 367-411). New York: Plenum.
- Ollendick, T. H., & Ollendick, D. G. (1997). General worry and anxiety in children. In *Session: Psychotherapy in Practice*, 3, 89-102.
- Ollendick, T. H., Yang, B., King, N. J., Dong, Q., & Akande, A. (1996). Fears in American, Australian, Chinese, and Nigerian children and adolescents: A cross-cultural study. *Journal of Child Psychology and Psychiatry*, 37, 213-220.
- Ollendick, T. H., Yule, W., & Ollier, K. (1991). Fears in British children and their relation to manifest anxiety and depression. *Journal of Child Psychology and Psychiatry*, 32, 321-331.
- Olweus, D. (1979). Stability of aggressive reaction patterns in males: A review. *Psychological Bulletin*, 86, 852-875.
- Patterson, G. R. (1976). The aggressive child: Victim and architect of a coercive system. In E. J. Mash, L. A. Hammerlynck, & L. C. Hardy (Eds.), *Behavior modification and families* (pp. 267-316). New York: Brunner/Mazel.
- Patterson, G. R. (1982). *Coercive family process*. Eugene, OR: Castalia.
- Patterson, G. R., Ray, R. S., Shaw, D. A., & Cobb, J. A. (1969). *Manual for coding family interaction* (6th ed.). Unpublished manuscript, University of Oregon.
- Peterson, D. R. (1961). Behavior problems of middle childhood. *Journal of Clinical and Consulting Psychology*, 25, 205-209.
- Pollard, S., Ward, E., & Barkley, R. A. (1983). The effects of parent training and Ritalin on the parent-child interactions of hyperactive boys. *Child and Family Behavior Therapy*, 5, 51-69.
- Prewitt-Diaz, J. (1989). *The process and procedures for identifying exceptional language minority children*. State College, PA: Pennsylvania State University.
- Prinz, R. (Ed.) (1986). *Advances in behavioral assessment of children and families*. Greenwich, CT: JAI Press.
- Quay, H. C. (1977). Measuring dimensions of deviant behavior: The Behavior Problem Checklist. *Journal of Abnormal Child Psychology*, 5, 277-287.
- Quay, H. C., & Peterson, D. R. (1967). *Manual for the Behavior Problem Checklist*. Champaign, IL: University of Illinois.
- Quay, H. C., & Peterson, D. R. (1975). *Manual for the Behavior Problem Checklist*. Unpublished manuscript.
- Quay, H. C., & Peterson, D. R. (1983). *Interim manual for the Revised Behavior Problem Checklist*. Unpublished manuscript, University of Miami.
- Rekers, G. A. (1984). Ethical issues in child behavioral assessment. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 244-262). Elmsford, NY: Pergamon.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Service.
- Reynolds, C. R., & Richmond, B. O. (1985). *Revised children's manifest anxiety scale manual*. Los Angeles: Western Psychological Services.
- Rutter, M. (1986). The developmental psychopathology of depression: Issues and perspectives. In M. Rutter, C. E. Izard, & P. B. Read (Eds.), *Depression in young people: Clinical and developmental perspectives* (pp. 3-30). New York: Guilford Press.
- Rutter, M., & Garmezy, N. (1983). Developmental psychopathology. In E. M. Hetherington (Ed.), *Socialization, personality, and social development: Vol 14. Mussen's Handbook of child psychology* (pp. 775-911). New York: Wiley.
- Scherer, M. W., & Nakamura, C. Y. (1968). A fear survey schedule for children (FSS-FC): A factor-analytic comparison with manifest anxiety (CMAS). *Behaviour Research and Therapy*, 6, 173-182.
- Schopler, E. (1974). Changes of direction with psychiatric children. In A. Davids (Ed.), *Child personality and*

- psychopathology: *Current topics* (Vol. I, pp. 205–236). New York: Wiley.
- Shaffer, D. (1992). *NIMH diagnostic interview schedule for children, Version 2.3*. New York: Columbia University Division of Child & Adolescent Psychiatry.
- Shapiro, E. S. (1984). Self-monitoring. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 148–165). Elmsford, NY: Pergamon.
- Shapiro, E. S., McGonigle, J. J., & Ollendick, T. H. (1980). An analysis of self-assessment and self-reinforcement in a self-managed token economy with mentally retarded children. *Journal of Applied Research in Mental Retardation*, 1, 227–240.
- Silverman, W. K., & Albano, A. M. (1996). *Anxiety Disorders Interview Schedule for DSM-IV*. San Antonio, TX: The Psychological Corporation.
- Silverman, W. K., & Nelles, W. B. (1988). The anxiety disorders interview schedule for children. *Journal of the American Academy of Child and Adolescent Psychiatry*, 27, 772–778.
- Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
- Smith, R. E., & Sharpe, T. M. (1970). Treatment of a school phobia with implosive therapy. *Journal of Consulting and Clinical Psychology*, 35, 239–243.
- Smucker, M. R., Craighead, W. E., Craighead, L. W., & Green, B. J. (1986). Normative and reliability data for the Children's Depression Inventory. *Journal of Abnormal Child Psychology*, 14, 25–39.
- Spielberger, C. D. (1973). *Preliminary manual for the State-Trait Anxiety Inventory for Children* ("how I feel questionnaire"). Palo Alto, CA: Consulting Psychologist Press.
- Sroufe, L. A., & Rutter, M. (1984). The domain of developmental psychopathology. *Child Development*, 55, 17–29.
- Staats, A. W. (1975). *Social behaviorism*. Homewood, IL: Dorsey Press.
- Staats, A. W. (1986). Behaviorism with a personality: The paradigmatic behavioral assessment approach. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 242–296). New York: Guilford Press.
- Stefanek, M. E., Ollendick, T. H., Baldock, W. P., Francis, G., & Yaeger, N. J. (1987). Self-statements in aggressive, withdrawn, and popular children. *Cognitive Therapy and Research*, 11, 229–239.
- Swann, G. E., & MacDonald, M. L. (1978). Behavior therapy in practice: A rational survey of behavior therapists. *Behavior Therapy*, 9, 799–807.
- Ullmann, L. P., & Krasner, L. (Eds.) (1965). *Case studies in behavior modification*. New York: Holt, Rinehart, & Winston.
- Vasquez Nuttall, E., DeLeon, B., & Del Valle, M. (1990). Best practice in considering cultural factors. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology II*, (pp. 219–233). Washington, DC: National Association of School Psychologists.
- Vasquez Nuttall, E., Sanchez, W., Borrás Osorio, L., Nuttall, R. L., & Varvogil, L. (1996). Assessing the culturally and linguistically different child with emotional and behavioral problems. In M. J. Breen & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/behavioral disorders: A handbook for school-based practitioners* (pp. 451–502). Austin, TX: ProEd.
- Wahler, R. G. (1976). Deviant child behavior in the family: Developmental speculations and behavior change strategies. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy* (pp. 516–543). Englewood Cliffs, NJ: Prentice-Hall.
- Wahler, R. G., House, A. E., & Stambaugh, E. E. (1976). *Ecological assessment of child problem behavior: A clinical package for home, school, and institutional settings*. Elmsford, NY: Pergamon.
- Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*, 3, 1–14.
- Winett, R. A., Riley, A. W., King, A. C., & Altman, D. G. (1989). Preventive strategies with children and families. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed., pp. 499–521). New York: Plenum.
- World Health Organization (1991). *International classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines* (10th ed.). Geneva, Switzerland: Author.
- Zatz, S., & Chassin, L. (1983). Cognitions of test-anxious children. *Journal of Consulting and Clinical Psychology*, 51, 526–534.

4.07

Principles and Practices of Behavioral Assessment with Adults

STEPHEN N. HAYNES

University of Hawaii at Manoa, Honolulu, HI, USA

4.07.1 INTRODUCTION	158
4.07.2 CLINICAL JUDGMENTS AND FUNCTIONAL ANALYSIS IN BEHAVIORAL ASSESSMENT	158
4.07.3 CONCEPTUAL FOUNDATIONS OF BEHAVIORAL ASSESSMENT	162
4.07.3.1 <i>Assumptions About the Causes of Behavior Disorders</i>	162
4.07.3.1.1 <i>Multiple causality</i>	163
4.07.3.1.2 <i>Multiple causal paths</i>	163
4.07.3.1.3 <i>Individual differences in causal variables and paths</i>	163
4.07.3.1.4 <i>Environmental causality and reciprocal determinism</i>	163
4.07.3.1.5 <i>Contemporaneous causal variables</i>	164
4.07.3.1.6 <i>Interactive and additive causality</i>	165
4.07.3.1.7 <i>Situations, setting events, and systems factors as causal variables</i>	165
4.07.3.1.8 <i>Dynamic causal relationships</i>	166
4.07.3.1.9 <i>Additional assessment implications of causal assumptions</i>	166
4.07.3.2 <i>Assumptions About the Characteristics of Behavior Problems</i>	166
4.07.3.2.1 <i>Behavior problems can have multiple response modes</i>	166
4.07.3.2.2 <i>Behavior problems have multiple parameters</i>	167
4.07.3.2.3 <i>Client can have multiple behavior problems</i>	168
4.07.3.2.4 <i>Behavior problems are conditional</i>	168
4.07.3.2.5 <i>The dynamic nature of behavior problems</i>	168
4.07.4 METHODOLOGICAL FOUNDATIONS OF BEHAVIORAL ASSESSMENT	169
4.07.4.1 <i>An Empirically Based Hypothesis Testing Approach to Assessment</i>	169
4.07.4.2 <i>An Individualized Approach to Assessment</i>	170
4.07.4.3 <i>Time-series Assessment Strategies</i>	170
4.07.4.4 <i>Quantitative and Qualitative Approaches to Behavioral Assessment</i>	171
4.07.5 BEHAVIORAL ASSESSMENT METHODS	171
4.07.5.1 <i>Behavioral Observation</i>	172
4.07.5.1.1 <i>Behavioral observation in the natural environment</i>	173
4.07.5.1.2 <i>Analogue observation</i>	175
4.07.5.2 <i>Self-monitoring</i>	176
4.07.5.3 <i>Psychophysiological Assessment</i>	177
4.07.5.4 <i>Self-report Methods in Behavioral Assessment</i>	178
4.07.5.5 <i>Psychometric Foundations of Behavioral Assessment</i>	179
4.07.6 BEHAVIORAL AND PERSONALITY ASSESSMENT	180
4.07.7 SUMMARY	181
4.07.8 REFERENCES	182

4.07.1 INTRODUCTION

Psychological assessment is the systematic evaluation of a person's behavior. The components of psychological assessment include the variables selected for measurement (e.g., beliefs, social behaviors), the measurement methods used (e.g., interviews, observation), the reduction and synthesis of derived data (e.g., whether summary scores are calculated for a questionnaire), and the inferences drawn from the data (e.g., inferences about treatment effectiveness). Psychological assessment affects the evolution of all social, cognitive, and behavioral science disciplines. The accuracy with which variables can be measured affects the degree to which relationships among behavioral, cognitive, environmental, and physiological events can be identified and explained. For example, our understanding of the impact of traumatic life stressors on immune system functioning, the relationship between depressed mood and self-efficacy beliefs, the effect of social response contingencies on self-injurious behavior, and the degree to which presleep worry moderates the impact of chronic pain on sleep quality depends on the strategies we use to measure these constructs.

Psychological assessment also affects clinical judgments. In adult treatment settings, clinical psychologists make judgments about a client's risk of suicide, whether treatment is warranted for a client, whether a client should be hospitalized, the variables that affect a client's behavior problems, and the best strategy for treating a client. Psychological assessment also helps the clinician select intervention goals and evaluate intervention effects.

There are many paradigms in psychological assessment. A psychological assessment paradigm is composed of a coherent set of assessment principles, values, assumptions, and methods of assessment. It includes assumptions about the relative importance of different types of behavior problems, the variables that most likely cause behavior problems, the most likely mechanisms through which causal variables operate, the importance and role of assessment in treatment design, and the best methods and strategies of assessment. A psychological assessment paradigm also includes guidelines for problem-solving, decision-making strategies, and data interpretation.

One powerful and evolving psychological assessment paradigm is behavioral assessment (Haynes & O'Brien, in press). Influenced by behavior-analytic, social-learning, and cognitive-behavioral therapy construct systems, the paradigm incorporates diverse methods of assessment but emphasizes naturalistic and

analogue observation, self-monitoring, and electrophysiological measurement. The behavioral assessment paradigm also has many methodological elements, including an emphasis on the use of minimally inferential constructs, time-series measurement, hypotheses testing, and an idiographic (i.e., focus on an individual client) approach to assessment.

This chapter focuses on clinical applications of behavioral assessment with adults. The chapter will outline the conceptual and methodological elements of behavioral assessment and indicate how they aid clinical judgment and decision-making with adult clients. To illustrate the underlying assumptions, methods, and strategies of behavioral assessment, the first section presents a functional analytic causal model of a client—a vector diagram of a behavioral clinical case conceptualization. Following a discussion of the principles and methods of behavioral assessment, subsequent sections briefly discuss the history of behavioral assessment and psychometric considerations. Developments in behavioral assessment and differences between behavioral and nonbehavioral assessment paradigms are also discussed.

4.07.2 CLINICAL JUDGMENTS AND FUNCTIONAL ANALYSIS IN BEHAVIORAL ASSESSMENT

One of the most important and complex clinical judgments is the clinical case conceptualization. The clinical case conceptualization is a metajudgment—it is composed of many lower-level judgments regarding a client's behavior problems and the factors that affect them. It is a synthesis of assessment- and research-based hypotheses about a client and its primary application is for designing the most effective treatment.

In behavioral assessment, clinical case conceptualization is often termed “functional analysis” (Haynes & O'Brien, 1990). (Terms with similar meanings include “clinical pathogenesis map” [Nezu & Nezu, 1989] and “behavioral case formulation” [Persons, 1989]. The term “functional analysis” is often used in applied behavior analysis to refer to the systematic manipulation of variables, in a controlled assessment setting, to determine their effect on designated dependent variables.) Functional analysis is a central component in the design of behavior therapy programs because of individual differences between clients—two clients can manifest the same primary behavior problem for different reasons and, consequently, receive different treatments. Behavioral interventions are often designed to

modify variables that are hypothesized to affect (i.e., account for variance in, trigger, maintain, moderate) problem behaviors and goals (Haynes, Spain, & Oliveira, 1993). Many permutations of causal variables can result in identical behavior problems, thereby resulting in different functional analyses and warranting different interventions.

Behavioral interventions are designed on the basis of many judgments about a patient reflected in a functional analysis. Clinical judgments with important implications for treatment decisions include the importance (e.g., severity, degree of risk associated with) of a client's multiple behavior problems; the relationships (e.g., strength, correlational vs. causal) among a client's multiple behavior problems, and the effects of behavior problems. Judgments about causal variables that affect the behavior problem (their importance, functional form, modifiability) are particularly important elements of the functional analysis.

There can be many errors in the clinical judgments that compose a clinical case conceptualization. Books by Eels (1997), Nezu and Nezu (1989), and Turk and Salovey (1988) discuss many of these errors. In brief, a clinician's judgments about a client can be affected by the clinician's preconceived beliefs, recent or particularly salient clinical experiences, selective attention to data that confirms the clinician's expectations, training-related biases, premature diagnoses, decisions based on initial impressions, and insufficient integrative abilities. These errors can reduce the validity of the case conceptualization and reduce the chance that the most effective treatment strategy will be selected for a patient.

The supraordinate goal of behavioral assessment is to reduce error and increase the validity in clinical judgments. The behavioral assessment paradigm suggests that clinical judgment error can be reduced to the degree that the assessor uses multiple sources of information, validated assessment instruments, time-series measurement strategies, focuses on multiple response modes, minimizes the inferential characteristics of variables, addresses behavior-environment interactions, and is guided by data from previously published studies (e.g., Persons & Fresco, 1998).

Haynes (1994), Haynes et al. (1993), Nezu and Nezu (1989), and Nezu et al. (1996) have outlined two methods to help the clinician systematically integrate the complex information contained in a functional analysis. These methods are the clinical pathogenesis map and functional analytic causal models (FACMs). Both involve systematic construction of vector diagrams of the component clinical judgments

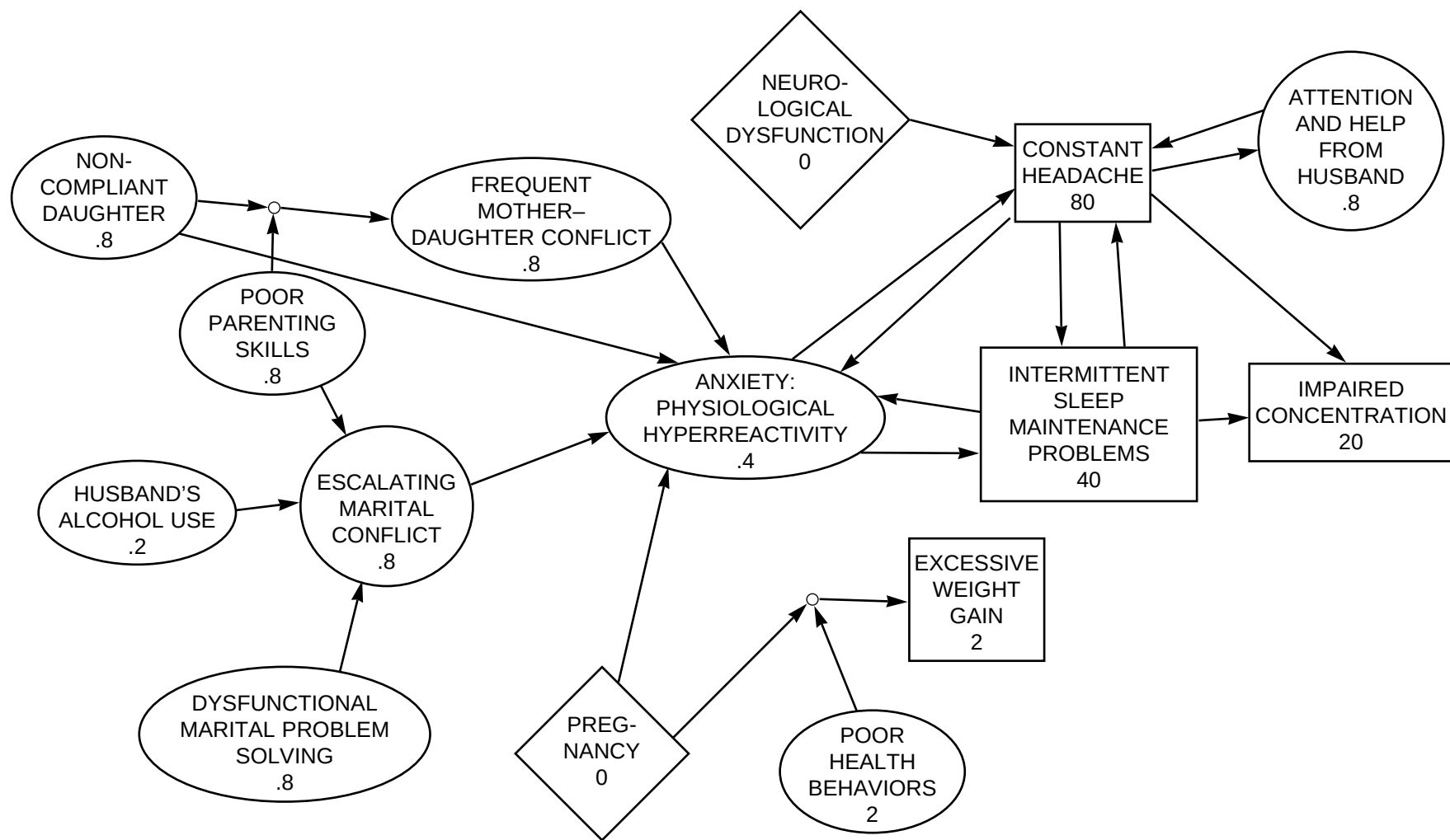
in a functional analysis and are designed to promote less intuitive intervention decisions. The clinical pathogenesis map and FACMs graphically model the clinician's hypotheses about a patient's behavior problems and goals and their relative importance, interrelationships, sequela and the strength, modifiability, and direction of action of causal variables. The FACM allows the clinician to estimate, quantitatively or qualitatively, the relative magnitude of effect of a particular treatment focus, given the clinician's hypotheses about the patient.

The FACM of a client presented in Figure 1 will be used to illustrate several underlying assumptions and methods of the behavioral assessment paradigm. The graphics in Figure 1 are explained in Figure 2. The client was a 35-year-old pregnant, married, unemployed woman (Mrs. A) who came to an outpatient mental health center complaining of constant headaches and sleeping problems. (This FACM was modified from one developed by Akiko Lau, University of Hawaii and discussed in Haynes, Leisen, & Blaine, 1997.) She was referred by her neurologist and had undergone extensive neurological and endocrinological examinations with negative results.

Mrs. A was first interviewed in an unstructured, open-ended manner, with the goal of encouraging her to talk about her behavior problems, goals, and the factors affecting them (Haynes, 1978). Structured interviews were then conducted to acquire more detailed information on specific behavior problems mentioned in the unstructured interview, such as her anxiety symptoms (Brown, DiNardo, & Barlow, 1994), marital relationship concerns (O'Leary, 1987), headache (James, Thorn, & Williams, 1993), sleep disturbance (Lichstein & Riedel, 1994), and other factors depicted in Figure 1. Validated questionnaires on marital satisfaction, child behavior problems, anxiety, and life stressors were also administered to provide additional information on issues raised in the interview process.

Because headaches and sleep difficulties were important problems, Mrs. A began daily self-monitoring after the second interview and continued to self-monitor throughout the assessment-treatment process. She recorded headache intensity and symptoms four times per day and each morning she recorded sleep-onset and awakenings for the previous night.

Marital conflict was a major concern for Mrs. A and one possible cause of her psychophysiological problems. Consequently, a one and a half hour assessment session (the third session) was conducted with her and her husband. During the session, the couple underwent a



Figures 1 and 2 An FACM of an outpatient woman with headaches and sleep disorders. The figures illustrate the relative importance of behavior problems, interrelationships among behavior problems, behavior problem sequela, casual relationships, and the modifiability of casual variables.

ILLUSTRATING A FUNCTIONAL ANALYSIS WITH A FUNCTIONAL ANALYTIC CAUSAL MODEL

IMPORTANCE/MODIFIABILITY OF VARIABLES
(using width of variable boundary and coefficients)



LOW IMPORTANCE/
MODIFIABILITY

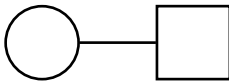


HIGH IMPORTANCE/
MODIFIABILITY

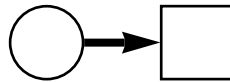
===== .4

TYPE AND DIRECTION OF RELATIONSHIP BETWEEN VARIABLES

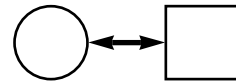
NONCAUSAL, CORRELATIONAL



UNIDIRECTIONAL CAUSAL



BIDIRECTIONAL CAUSAL



=====

SYMBOLS



ORIGINAL,
UNMODIFIABLE
CAUSAL
VARIABLE



CAUSAL
VARIABLE;
MEDIATING
VARIABLE

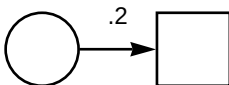


BEHAVIOR PROBLEM;
EFFECT OF BEHAVIOR
PROBLEM

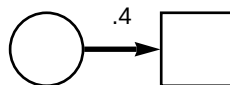
=====

STRENGTH OF RELATIONSHIP BETWEEN VARIABLES

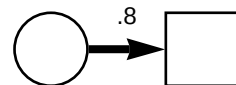
INDICATED BY ARROW THICKNESS; MORE PRECISELY BY COEFFICIENTS



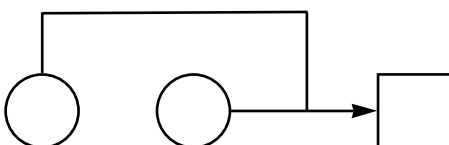
WEAK



MODERATE



STRONG



MEDIATING
RELATIONSHIP

conjoint structured interview about their marital relationship (e.g., perceived conflicts, strengths, spousal excesses and deficits, marital goals), and Mr. A also completed a marital satisfaction questionnaire. The couple participated in an analogue communication assessment, in which they discussed for 10 minutes their conflicts regarding disciplining their 12-year-old daughter. The conversation was recorded and later coded by the assessor.

Based on interview and questionnaire data, conflicts with her daughter were another source of distress for Mrs. A. A joint assessment session (the fourth session) was conducted in which the daughter was interviewed about her perception of family issues. Also, the mother and daughter were observed in two structured assessment settings: while trying to resolve one of their frequent sources of conflict (the daughter's refusal to do her school work) and while Mrs. A attempted to help her daughter with some school work.

The functional analytic causal model of Mrs. A emphasizes many elements of a clinical case conceptualization that are important for behavioral treatment decisions. (Many other factors—e.g., treatment history, client cognitive resources, cost-efficiency of treatments, responses of persons in the client's environment—affect treatment design in addition to those included in a FACM.) Important and controllable functional relationships are highlighted in the FACM because of their clinical utility. The FACM for Mrs. A recognizes some unmodifiable variables but these have no clinical utility. Unidirectional and bidirectional causal relationships are shown because they can significantly affect decisions about what variables should be targeted in treatment. Treatment decisions are also affected by the strength of causal relationships and the degree of modifiability of causal variables, depicted in Figure 1.

Before considering the specific assumptions of the behavioral assessment paradigm that influenced the assessment strategy outlined above and the clinical judgments summarized in Figure 1, several additional attributes of the functional analysis should be briefly noted. First, the functional analysis (and the FACM) reflects the clinician's current judgments about a client. It is a subjectively derived (although data-influenced), hypothesized model. It is also unstable, in that it can change with changes in naturally occurring causal variables, with the acquisition of additional data, and as a result of treatment. For example, a change in the variables that affected Mr. A's drinking could lead to a significant change in the FACM for Mrs. A. A FACM for a client may be

conditional. For example, some variables affecting Mrs. A's behavior problems may change after the birth of her child.

A final note on the limitations of the functional analysis. Despite its central role in behavior therapy, the functional analysis is limited in several ways: (i) the best assessment methods for developing a functional analysis have not been identified, (ii) the best methods for formulating a functional analysis from assessment data have not been determined, and (iii) for many behavior problems, the incremental utility and cost-effectiveness of the functional analysis have yet to be established.

4.07.3 CONCEPTUAL FOUNDATIONS OF BEHAVIORAL ASSESSMENT

Many methodological elements of the behavioral assessment paradigm, such as the preferred methods of assessment and the variables targeted in assessment, are influenced by its underlying assumptions. The following sections review two sets of assumptions: (i) those concerning the causal factors associated with behavior problems and goals, and (ii) those concerning the characteristics of behavior problems. This section also discusses implications of these assumptions for behavioral assessment strategies with adults. More extensive discussions of underlying assumptions in behavioral assessment can be found in Bandura (1969), Barrios (1988), Bellack and Hersen, (1988), Bornstein, Bornstein, and Dawson (1984), Ciminero, Calhoun, and Adams (1986), Cone (1988), Eysenck, (1986), Haynes (1978), Hersen and Bellack (1998), Johnston and Pennypacker, (1993), Kratochwill and Shapiro (1988), Mash and Terdal (1988), Nelson and Hayes (1986), O'Donohue and Krasner, (1995), Ollendick and Hersen (1984, 1993), Strosahl and Linehan (1986), and Tryon (1985).

4.07.3.1 Assumptions About the Causes of Behavior Disorders

Psychological assessment paradigms differ in the assumptions they make regarding the causes of behavior disorders. Although causal assumptions and the identification of causal variables in pretreatment assessment are less important for treatment paradigms with limited treatment options (e.g., Gestalt, transactional therapies), the identification of potential causal variables is a primary objective in pretreatment behavioral assessment. This is because hypothesized controlling variables are targeted for modification in behavior therapy and it is

presumed that causal variables may vary across patients with the same behavior problems. The variables presumed to cause Mrs. A's sleep problems may not operate for other patients with identical sleep problems. Consequently, other patients with the same sleep disorder would be treated differently.

Behavioral assessment strategies are guided by several empirically based and interrelated assumptions about the causes of behavior problems. These assumptions include: multiple causality, multiple causal paths, individual differences in causal variables and paths, environmental causality and reciprocal determinism, contemporaneous causal variables, the dynamic nature of causal relationships, the operation of moderating and mediating variables, interactive and additive causality; and situations, setting events, and systems factors as causal variables and dynamical causal relationships.

4.07.3.1.1 Multiple causality

Behavior problems are often the result of multiple causal variables acting additively or interactively (Haynes, 1992; Kazdin & Kagan, 1994). This is illustrated in Figure 1 by the multiple factors influencing Mrs. A's headaches. Although some behavior problems and the behavior problems of some individuals (e.g., asthma episodes that are mostly triggered by exposure to specific allergens) may primarily be the result of single causal variables, multivariate causal models have been proposed for most adult behavior disorders, including schizophrenia, chronic pain, sleep disorders, paranoia, personality disorders, child abuse, and many other behavior disorders (see reviews in Gatchel & Blanchard, 1993; Sutker & Adams, 1993).

4.07.3.1.2 Multiple causal paths

A causal variable may also affect a behavior problem through multiple paths. Note that for Mrs. A, physiological hyperreactivity can directly influence sleep but hyperreactivity can also influence sleep because it produces headaches. Similarly, there may be many paths through which social isolation increases the risk of depression (e.g., by restricting the potential sources of social reinforcement, by increasing dependency on reinforcement from a few persons) and many paths through which immune system functioning can be impaired by chronic life stressors (e.g., dietary changes, reduction of lymphocyte levels).

4.07.3.1.3 Individual differences in causal variables and paths

Models of causality for behavior problems are further complicated because the permutations of causal variables and causal mechanisms can differ across clients with the same behavior problem. For example, there can be important differences in the causal variables and causal paths among persons reporting chronic pain (Waddell & Turk, 1992), exhibiting self-injurious behaviors (Iwata et al., 1994), or who complain of difficulties in initiating and maintaining sleep (Lichstein & Riedel, 1994). Some differences in causality may covary with dimensions of individual differences. For example, the causes of depression, marital distress, and anxiety may vary as a function of ethnicity, age, gender, religion, sexual orientation, and economic status (see discussions in Marsella & Kameoka, 1989).

4.07.3.1.4 Environmental causality and reciprocal determinism

The behavioral assessment paradigm also stresses the importance of environmental causality and behavior–environment interactions (McFall & McDonel, 1986). Many studies have shown that it is possible to account for variance in many behavior problems by examining variance in response contingencies (e.g., how others respond to self-injurious behaviors, depressive statements, or asthma episodes can affect the parameters of those behaviors—a “parameter” of a behavior refers to a quantitative dimension, such as rate, duration, magnitude, and cyclicity), situational and antecedent stimulus factors (e.g., alcohol use may vary reliably across social settings, anxiety episodes may be more likely in more interpersonally stressful environments), and other learning principles (e.g., modeling, stimulus pairings; see discussions in Eysenck & Martin, 1987; O'Donohue & Krasner, 1995).

An important element of the principle of environmental causality is reciprocal determinism, (i.e., bidirectional causality, reciprocal causation; Bandura, 1981)—the idea that two variables can affect each other. In a clinical context, reciprocal determinism refers to the assumption that clients can behave in ways that affect their environment which, in turn, affects their behavior. For example, a client's depressive behaviors (e.g., reduced social initiations and positive talk) may result in the withdrawal of the client's friends, increasing the client's loss of social reinforcement and increasing the client's depressive mood and behaviors. A hospitalized paranoid patient may behave

suspiciously with staff and other patients. These behaviors may cause others to avoid the patient, talk about him/her and behave in many other ways that confirm and strengthen the patient's paranoid thoughts. With Mrs. A, we presume that there are some ways that Mrs. A is behaving that might contribute to her marital distress and difficulties. An emphasis on bidirectional causation does not negate the possibility of unidirectional environmental causal factors. In some distressed marriages, for example, a spouse may independently be contributing to marital distress by being verbally abusive or unsupportive. However, pure unidirectional causation may be rare.

Viewing clients within a reciprocal determinism framework effects the focus of assessment and treatment. Clients are considered active participants in their own lives, as active contributors to their goal attainment and to their behavior problems. Consequently, clients are encouraged to participate actively in the assessment–treatment process. Assessment goals include identifying the ways that clients may be contributing to their behavior problems and ways they can contribute to the attainment of treatment goals.

One consequence of reciprocal determinism is that labels such as “behavior problem (dependent variable)” or “causal variable (independent variable)” become less distinct. Often, either variable in a bidirectional relationship can be described as a behavior problem and a causal variable—each variable can be either or both. Which variables are described as problem vs. cause depends more on convention or the intent of the assessor and client than on the characteristic of the functional relationships. As indicated in the functional analytic causal models (e.g., Haynes, 1994), treatment decisions are dictated more by estimates of the strength of causal relationships than by the label of the variable.

The concept of reciprocal determinism also promotes a behavioral skills focus in assessment and treatment. A client's behavior problems are presumed to be a partial function of their behavioral repertoire. Their behavioral excesses, deficits, and strengths are presumed to affect whether they will experience problems in some situations, the type and magnitude of behavior problem experienced and how long the problem persists. For example, a behavior skills assessment with a socially anxious client might focus on specific deficits that prevent the client from forming more frequent and satisfying friendships. Similar to a task analysis, the necessary skills for attaining a treatment goal (e.g., establishing positive friendships) are broken down into molecular components and the client's abilities on these components are

evaluated. With Mrs. A, it would be important to determine what additional parenting and marital communication skills might help Mrs. A develop a more positive relationship with her daughter and husband. An example would include the ability to clearly and positively talk about her ideas and concerns.

Cognitive skills are often targeted by behavioral assessors. The clients' beliefs, expectancies, deductions, and other thoughts regarding their capabilities in specific situations (e.g., Linscott & DiGiuseppe, 1998) are often considered essential elements for effective functioning. A molar-level skill is adaptive flexibility—an overarching goal of behavior therapy is to help the client to develop behavior repertoires that facilitate adaptability to various, novel and changing environments.

4.07.3.1.5 Contemporaneous causal variables

The behavioral assessment paradigm emphasizes the relative importance and utility of contemporaneous rather than historical, causal factors. It is presumed that a more clinically useful, and sometimes more important, source of variance in a client's behavior problems can be identified by examining the client's current, rather than historical, learning experiences, social contingencies, and thoughts. For example, suspicious thoughts and behaviors can undoubtedly be learned as a child from parental models (e.g., parents who teach a child to be mistrustful of others' intentions; Haynes, 1986). However, early parent–child learning experiences are difficult to identify and “treat” in therapy. Assessment efforts might more profitably be focused on contemporaneous causal variables for paranoid thoughts—such as restricted social network that precludes corrective feedback about misperceptions, social skills deficits, hypersensitivity to negative stimuli or negative scanning, or failure to consider alternative explanations for ambiguous events. These can also be important causal variables for a client's paranoid behaviors and are more amenable than historical events to intervention.

The emphasis on contemporaneous, reciprocal, behavior–environment interactions dictates an emphasis on particular methods of assessment. For example, naturalistic observation, analogue observation, and self-monitoring are better suited than retrospective questionnaires to measuring contemporaneous, reciprocal dyadic interactions. Additionally, in behaviorally oriented interviews and questionnaires clients are more often asked about current than about past behavior–environment interactions (Jensen & Haynes, 1986; Sarwer & Sayers, 1998).

An emphasis on contemporaneous reciprocal determinism is compatible with the causal role of genetic and physiological factors, and early learning experiences. Evidence from many sources suggests that genetic factors, neuro-physiological mechanisms, medical disorders, and early learning (e.g., early traumatic experiences) can serve as important causal variables for behavior problems (see reviews in Asteria, 1985; Haynes, 1992; Sutker & Adams, 1993). Sometimes, physiological, behavioral, and cognitive variables are different modal expressions of the same phenomena.

The emphasis on contemporaneous behavior and environmental causality is evident in the contemporaneous focus of many behavioral assessment interviews. However, behavioral assessors differ in their emphasis on historical data. Joseph Wolpe, for example, emphasized the importance of gathering a complete clinical history for patients before therapy (Wolpe & Turkat, 1985). Often historical information can aid in the development of hypotheses regarding the time-course and causes of behavior problems. For example, a careful interview about past behaviors, events and treatment experiences can help determine if Mrs. A may be experiencing neurological deficits (e.g., she had a minor head injury two years prior to this assessment) and may help estimate the degree to which her health-related behaviors (e.g., poor diet and exercise habits) are modifiable.

4.07.3.1.6 Interactive and additive causality

In the section on multiple causality, I noted that a behavior problem often results from multiple causal factors acting concurrently—this is an additive model of causality. Causal variables can also interact—this is a multiplicative model of causality. Interactive causality occurs when the causal effects of one variable vary as a function of the values of another causal variable (see discussion in Haynes, 1992). Furthermore, the effects of the variables in combination often cannot be predicted by simply summing their independent effects. A longitudinal study by Schlundt, Johnson, and Jarrell (1986) demonstrated interactive causal effects with bulimic clients. The probability of postmeal purging was significantly related to a history of recent purges (i.e., purging tended to occur in cycles). However, the social context within which eating occurred affected the strength of the relationship between those two variables. The chance of purging was higher when the person had recently purged, but especially higher if the person ate alone. The effect of each causal

variable (purging history, social context) depended on the value of the other causal variable.

Diathesis-stress models of psychopathology are common exemplars of interactive causality (e.g., Barnett & Gotlib, 1988). Diathesis-stress models suggest that environmental stressors and physiological or genetic vulnerability (or genetic and later physiological challenges) interact to affect the probability that a particular behavior disorder will occur.

4.07.3.1.7 Situations, setting events, and systems factors as causal variables

One assumption of the behavioral assessment paradigm is that the probability (or another parameter such as magnitude) of behavior problems varies across situations, settings, and antecedent stimuli (e.g., discrete and compound antecedent stimuli, contexts, discriminative stimuli for differential reinforcement contingencies); behavior problems are conditional. The conditional nature of behavior problems has important causal implications because it marks the differential operation of causal factors. Mrs. A was more likely to experience anxiety symptoms in the presence than in the absence of her daughter. The presence of the daughter marked the operation of a causal relationship and suggested to the assessor that the mother–daughter interactions should be explored in greater detail.

A situational model of behavior problems contrasts with traditional personality trait models, which emphasize a higher degree of cross-situational consistency of behavior and some enduring trait of the person as the primary causal variable for behavior problems (see subsequent discussion of personality assessment). However, situational and trait models of behavior problems can be compatible. Knowledge of the robust behaviors of a client (e.g., those behaviors that do not vary to an important degree across conditions) and knowledge of the situational factors that influence the client's behavior can both contribute to a functional analysis of the client. This “interactional” perspective is a welcomed refinement of exclusively trait models (see discussions by Mischel, 1968; McFall & McDonel, 1986).

Cross-situational consistency can vary across different behaviors, individuals, and situations; relative cross-situational consistency in behavior can occur, but may not. Because the assessor does not have prior knowledge of the degree of cross-situational consistency of a client's behavior problems, the assessor must evaluate their conditional nature. Unfortunately, strategies and classification schema for situations have not been developed.

Although the behavioral assessment paradigm emphasizes contemporaneous causal factors (e.g., a SORC [stimulus, organism, response, contingency] model; Goldfried, 1982), extended social systems can play an important causal role. Mrs. A's marital satisfaction may be affected by her relationships with her friends and family. Her daughter's behavior problems in the home may be affected by the social and academic environment of her school. Assessment efforts cannot be confined to individual elements extracted from a complex array of interacting variables. Chaos theory and dynamic modeling also suggest that it may be difficult to develop powerful predictive or explanatory models unless we measure behavior within the complex dynamical systems in which the behavior is imbedded (Vallacher & Nowack, 1994).

4.07.3.1.8 Dynamic causal relationships

All elements of a functional analytic causal model—the causal variables that affect a client's behavior problems, the strengths of causal relationships, moderating variables, for example, are nonstationary (Haynes, Blaine, & Meyer, 1995). Causal relationships for a client can be expected to change across time in several ways. First, new causal variables may appear: Mr. or Mrs. A may develop new health problems; Mr. A may lose his job. Second, a causal variable may disappear: Mr. A may stop drinking; Mrs. A may give birth to her baby. Third, the strength and form of a causal relationship are likely to change over time. There may be a decrease in sleep disruption originally triggered by a traumatic event; marital distress that originally caused depressive symptoms may be exacerbated by continued depressive reactions. Fourth, moderating variables may change: Clients may change their expectancies about the beneficial effects of alcohol (Smith, 1994). In causal models of behavior disorders, a moderating variable is one that changes the relationship between two other variables. For example, "social support" would be a moderating variable if it affected the probability that an environmental disaster would be associated with symptoms of post-traumatic stress disorder (PTSD).

4.07.3.1.9 Additional assessment implications of causal assumptions

Emphases on multivariate, idiosyncratic, interactive, reciprocal deterministic, and dynamic causal models have several implications for behavioral assessment strategies that were

briefly noted in the previous sections and will be discussed in greater detail later in this chapter. The assessment implications include: (i) pretreatment assessment should be broadly focused on multiple variables; (ii) classification will usually be insufficient to identify the causal variables operating for a particular client and insufficiently precise to identify the client's behavior problems; (iii) assessors should avoid "premature" or "biased" presumptions of causal relationships for a client and draw data-based inferences whenever possible; (iv) a valid functional analysis is most likely to result from assessment that focuses on multiple response modes, uses multiple methods; and gathers data from multiple sources; (v) it is important to identify the mechanisms underlying causal relationships (see also, discussion in Shadish, 1996); and (vi) a time-series assessment strategy can be an effective method of identifying and tracking behavior problems and potential causal factors.

4.07.3.2 Assumptions About the Characteristics of Behavior Problems

Behavioral assessment strategies and the resulting clinical case conceptualizations are strongly affected by assumptions of the behavioral assessment paradigm about the characteristics of behavior problems. Several of these assumptions were mentioned earlier in this chapter. They include the multimodal and multiparameter characteristics of behavior problems, differences among clients in the importance of behavior problem modes and parameters, the complex interrelationships among a client's multiple behavior problems, and the conditional and dynamic natures of behavior problems.

4.07.3.2.1 Behavior problems can have multiple response modes

Adult behavior problems can involve more than one response mode. For example, PTSD can involve physiological hyperreactivity, subjective distress, avoidance of trauma-related situations and thoughts, and distressing recollections and dreams of the traumatic event (American Psychiatric Association, 1994; Kubany, 1994). The degree of covariation among modes of a behavior problems can vary across persons (note that the *Diagnostic and statistical manual of mental disorders* [4th ed.] requires that a client manifest only one of five major [category B] symptoms for a diagnosis of PTSD). For

example, some PTSD clients show only slight evidence of distressing recollections of the event but show significant physiological reactivity to event-related cues while others show the opposite pattern.

Low levels of covariation among the multiple modes of behavior problems have been found in both group nomothetic and single-subject time-series research (see discussion in Gannon & Haynes, 1987). Acknowledging that the apparent magnitude of covariation reflects the ways in which the modes are measured, these findings suggest that behavior problem modes can be discordant across persons and for one person across time. (See discussions in Cone (1979) and Eifert and Wilson (1991). Different response modes are often measured with different methods. Different response modes can also have different response latencies which can reduce apparent magnitudes of covariation if all modes are sampled simultaneously.)

Discordance among response modes for some clients has many clinical implications. Different response modes of a behavior problem may have different latencies in response to treatment. Furthermore, some treatments may have a stronger effect on some modes than on others. Consequently, the effects of treatment may be judged differently depending on which response mode is measured by which method. More important for a functional analysis of a patient, different response modes can be affected by different causal factors and respond differently to the same causal factor. Therefore, the selection of the most important mode for a client's behavior problem can affect the functional analysis and the intervention program designed for that client.

Assessment strategies should be congruent with the multimodal nature of behavior problems. First, because causal inferences are an important component of the functional analysis and guide treatment decisions, the primary mode of a client's behavior problem should be identified. Second, as the prior example of PTSD illustrated, diagnosis may be helpful but is usually an insufficient basis for the identification of the most important response modes for a client. Third, inferences regarding treatment effects for one mode may not be generalizable to other modes. In sum, behavioral assessment should have a multimodal focus.

4.07.3.2.2 Behavior problems have multiple parameters

As previously mentioned, each behavior problem mode can have multiple parameters. Parameters are quantitative dimensions, such as

duration, magnitude, and rate, that can be used to describe a behavior. As with response modes, there are important between-client differences in the relative importance of different behavior problem parameters. For example, Mrs. A reported mildly intense but constant headaches and intermittent but severe sleep disruption. Other clients could report the same problems with different parameters (e.g., infrequent but debilitatingly intense headaches). Similarly, some clients report frequent but severe episodes of depression that last for only a few days; others report episodes of mild to moderate depression that can last months.

Multiple parameters of behavior disorders have important implications for the functional analysis because different parameters may be affected by different causal variables. For example, Barnett and Gotlib (1988) reviewed the literature on depression and suggested that learned helplessness beliefs seem to affect the duration and magnitude of depressive behaviors. However, learned helplessness beliefs could not account for the onset of depressive behaviors. Consequently, the functional analysis and treatment of a client with frequent depressive episodes might be different from the functional analysis and treatment of a client with infrequent but persistent depressive episodes.

One assessment implication that permeates many assumptions underlying behavioral assessment is that aggregated measures are insufficiently precise for a functional analysis. Between-person differences in the importance of behavior problem modes and parameters mandate careful specification and measurement of multiple modes and parameters. Measures of behavior problems that aggregate across modes, parameters, situations, for example, a single measure of "depression" or "anxiety," will often be insufficiently precise for the development of a functional analysis and for the design of intervention programs. Aggregated indices may be helpful as a general index of the magnitude or generality (or unconditional probability) of a behavior problem, but are insufficient for the precise inferences that are necessary for the evolution of assessment paradigms, functional analyses, and treatment strategies.

One helpful strategy for gathering more precise data on behavior problem parameters is the construction of a time-course for behavior problem—a time line reflecting occurrence, magnitude, and duration of behavior problems. An example of this method is the "timeline followback" by Sobell, Toneatto, and Sobell (1994) to establish a time-course of substance use.

4.07.3.2.3 Client can have multiple behavior problems

Many clients have multiple behavior problems. For example, Beck and Zebb (1994) reported that 65–88% of panic disorder patients have a coexisting behavior disorder, Figley (1979) reported a high incidence of comorbidity for PTSD, Regier et al., (1990) noted that drug users often have other concurrent behavior problems. Similar observations of comorbidity have been reported for panic disorders (Craske & Waikar, 1994) and depression (Persons & Fresco, 1996).

A client with multiple behavior problems challenges the clinical judgment capabilities of the assessor and complicates the functional analysis because the mode and parameter of each behavior problem can be affected by multiple causal variables: functional analytic causal models were developed partly as a method of organizing these clinical judgments. Additionally, multiple behavior problems can have complex causal and noncausal relationships. Note the relationships between sleep and headache problems for Mrs. A illustrated in Figure 1. Beach, Sandeen, and O'Leary (1990) observed a reciprocal causal relationship between marital distress depression for some patients (with many variables moderating that relationship). Hatch (1993) observed that depression can affect pain perception of headache patients and that headaches may contribute to the occurrence of depressive episodes for some patients.

The assumption that clients may have more than one behavior problem has several implications for behavioral assessment strategies. First, initial assessment, such as the intake interview (Kerns, 1994; Sarwer & Sayers, 1998) must be broadly focused to identify a client's multiple behavior problems. Premature narrowing of the assessment focus may preclude the identification of important behavior problems. Following a broad survey, subsequent assessment efforts can be focused on problem specification and the functional relationships relevant to each behavior problem. The functional analysis and intervention decisions will also be affected by estimates of the form and strength of relationship among, and relative importance of, a client's behavior problems. Multiple behavior problems also mandate a multivariate focus in treatment outcome evaluation.

Sometimes, the identification of functional response groups will aid in treatment decisions (Sprague & Horner, 1992). A functional response group is a set of behaviors, which may differ in form, that are under the control of the same contingencies (a set of behaviors that

has the same function, (Haynes, 1996). Adaptive elements of the response class can sometimes be strengthened to weaken maladaptive elements of that class. Relaxation skills may be taught as a substitute for dysfunctional ways of reducing physiological hyperarousal. Effective communication skills may reduce the frequency of self-injurious behavior for some developmentally disabled individuals (Durand & Carr, 1991).

4.07.3.2.4 Behavior problems are conditional

As noted earlier in this chapter, behavior problems seldom occur randomly or unconditionally. Although it is difficult to predict the occurrence of many behavior problems, the probability of occurrence often varies as a function of settings, antecedent stimuli, environmental contexts, physiological states, and other discriminative stimuli (Gatchel, 1993; Glass, 1993; Ollendick & Hersen, 1993). It was previously noted that identifying sources of variance in behavior problems can help the assessor to identify causal variables and mechanisms. For example, behavioral assessors attempt to identify the conditions that trigger the startle responses of a client with PTSD (Foa, et al., 1989), that triggers a client's asthma episodes (Creer & Bender, 1993), and the conditions associated with marital violence (O'Leary, Vivian, & Malone, 1992) to develop a functional analysis and design the most appropriate intervention program.

The conditional nature of behavior problems further diminishes the clinical utility of aggregated measures of a behavior problem—assessment instruments that provide a “score” without providing more precise information of the conditional nature of the behavior problem. Assessment instruments should help the assessor examine the conditional probabilities of the behavior problem or the magnitude of shared variance between the behavior problem and multiple situational factors. For Mrs. A, the assessor would try to determine the situations that provoked conflict with her daughter, and to determine the events that increased or decreased the intensity of her headaches. Behavioral questionnaires and interviews, self-monitoring, and naturalistic observation can aid in identifying the conditional aspects of behavior problems.

4.07.3.2.5 The dynamic nature of behavior problems

The parameters and qualitative aspects (e.g., topography, form, characteristics) of behavior problems can change over time. The frequency,

intensity, and content of arguments between Mr. and Mrs. A will probably change within a few weeks and months. The magnitude, frequency, duration, and form of a clients' PTSD symptoms, paranoid delusions, nightmares, excessive dieting, and pain can change in important ways across time. Also, new behavior problems may occur and some behavior problems may become less important.

Dynamic behavior problems and other variables can only be sensitively tracked by measuring them frequently, using time-series assessment strategies. The frequency with which dynamic variables should be sampled depends on their rate of change. Collins and Horn (1991), Heatherton and Weinberger (1994), Johnston and Pennypacker (1993), Kazdin (1997), and Kratochwill and Levin (1992) discuss instruments, strategies, and issues in the measurement of dynamic variables. Frequent measurement of behavior problems can also help the assessor identify important causal relationships. For example, recent changes in the magnitude of a client's depressive symptoms may provide cues about environmental or cognitive causal factors. Changes in Mrs. A's sleep patterns could trigger inquiries about events preceding the change. Self-monitoring, brief structured interviews and short questionnaires are particularly amenable to time-series assessment.

4.07.4 METHODOLOGICAL FOUNDATIONS OF BEHAVIORAL ASSESSMENT

The methodological elements of the behavioral assessment paradigm, the preferred strategies of behavioral assessment, are dictated by the assumptions about behavior and its causes outlined in the previous sections. Many of these methodological elements were introduced earlier in this chapter and are outlined in Table 1.

This section will discuss three of the elements of the behavioral assessment paradigm delineated in Table 1: (i) the emphasis on empirical hypothesis-testing, (ii) the idiographic emphasis, and (ii) the use of time-series assessment strategies. These and other methodological elements from Table 1 have been presented in Cone and Hawkins (1977), Haynes (1978), Hersen and Bellack (1996), Johnston and Pennypacker, (1993), Mash and Terdal (1988), Nelson and Hayes (1986), Ollendick and Hersen (1993), and Shapiro and Kratochwill (1988).

Most of the methodological elements of the behavioral assessment paradigm, particularly

the assessment strategies, serve the priority placed on a scholarly, empirically based approach to clinical assessment. It is assumed that clinically useful knowledge about behavior problems is best acquired through the frequent administration of carefully constructed assessment instruments that are precisely focused of multiple, minimally inferential variables. Behavioral assessors are likely to eschew assessment instruments that are poorly validated, provide indices of vaguely defined and highly inferential constructs, and have unsubstantiated clinical utility.

4.07.4.1 An Empirically Based Hypothesis Testing Approach to Assessment

A hypothesis testing and refinement climate guides the behavioral assessment of adults. The assessor makes many tentative judgments about the client beginning early in the preintervention assessment phase. The clinician estimates the relative importance of the client's behavior problems and goals, the variables that influence problems and goals, and other elements of the functional analysis. Based on these early clinical judgments, the assessor also begins to estimate the most effective methods of intervention and selects additional assessment instruments, (see discussions in Eels, 1996; Haynes et al., 1997; Nezu et al., 1996; O'Brien & Haynes, 1995; Persons & Bertagnolli, 1994; Turk & Salovey, 1988). These hypotheses are tested and refined as assessment continues. With Mrs. A, for example, initial judgments that deficits in marital communications skills were functionally related to their marital conflicts would be evaluated subsequently through analogue communication assessment and by teaching the couple more positive discussion strategies. If their communication skills increased but their marital conflicts did not decrease invalid hypotheses may have been initially drawn about the causes of this couple's marital conflicts. There are other possible explanations for a lack of observed covariation in a causal relationship. For example, if another causal factor or moderating variable became operational while communication skills were strengthened, there could appear to be no causal relationship between communication skills and conflict (Shadish, 1996).

A scholarly hypotheses-testing approach to psychological assessment requires that the results of assessment and contingent clinical inferences be viewed skeptically. Clinical judgments are always based on imperfect measurements of nonstationary data and are intrinsically subjective. The validity and utility

Table 1 Methodological emphases of the behavioral assessment paradigm.*Assessment strategies*

1. Idiographic assessment; a focus on the client's specific goals, individual behavior problems; individually tailored assessment; a de-emphasis on normatively based assessment (e.g., trait-focused questionnaires).
2. A hypothesis-testing approach to assessment and treatment (including the use of interrupted time-series designs).
3. Time-series assessment strategies (as opposed to single-point or pre-post measurement strategies); frequent measures of important variables.
4. Multimethod and multi-informant measurement of variables across multiple situations.
5. Quantification of variables (measurement of rates, magnitudes, durations).
6. The use of validated assessment instruments in conditions in which sources of measurement error are minimized.

The focus of assessment

7. Precisely specified, lower-level, and less inferential constructs and variables.
8. Observable behavior (as opposed to hypothesized intrapsychic events).
9. Client-environment interactions, sequences, chains, and reciprocal interactions.
10. Behavior in the natural environment.
11. Events that are temporally contiguous to behavior problems.
12. Multiple client and environmental variables in pretreatment assessment (multiple behavior problems, causal variables, moderating and mediating variables).
13. Multiple targets in treatment outcome evaluation (e.g., main treatment effects, side effects, setting and response generalization).
14. Multiple modes and parameters of behavior and other events.
15. Extended systems (e.g., family, work, economic variables).

Source: Haynes (1996b).

These are relative emphases whose importance varies across behavioral assessment subparadigms (e.g., behavior analysis, cognitive-behavioral). Many are compatible with other psychological assessment paradigms (e.g., neuropsychological, educational achievement).

of clinical judgments will covary with the degree to which they are guided by the assessment principles outlined in Table 1. The assessor can also reduce some of the biases in clinical judgments by basing them on assessment data and being receptive to clinical assessment data that is inconsistent with those judgments.

4.07.4.2 An Individualized Approach to Assessment

Partially due to between person differences in behavior problems, goals, and functional relationships, the behavioral assessment paradigm emphasizes individualized assessment—an idiographic approach to assessment (Cone, 1986). An individualized approach to assessment is manifested in several ways: (i) self-monitoring targets and sampling procedures are often tailored to the individual client; (ii) role play scenarios and other assessment instruments are often individually tailored (e.g., Chadwick, Lowe, Horne, & Higson, 1994); (iii) client-referenced and criterion-referenced assessment instruments, in contrast to norm-referenced assessment instruments, are often used; (iv) treatment goals and strategies are often individually tailored (e.g., de Beurs, Van Dyck, van Balkom, Lange, & Koele, 1994), and (v) within-subject, interrupted time-series and mul-

tivariate time-series designs are often used in research.

4.07.4.3 Time-series Assessment Strategies

As indicated in previous sections, time-series measurement of independent and dependent variables across time (e.g., the frequent [e.g., >40] daily samples of a client's behavior problems and causal variables) is an important strategy in behavioral assessment. It is a powerful strategy in clinical research and has many advantages in clinical assessment. First, it can help estimate causal and noncausal relationships for a participant's behavior problems. By subjecting the data to sequential analyses or cross-lagged correlations time-series assessment can provide information on conditional probabilities and suggest possible causal mechanisms for behavior problems (Bakeman & Gottman, 1986; Moran, Dumas, & Symons, 1992; Tryon, 1998).

A major advantage is that time-series assessment allows the researcher and clinician to control for and examine the temporal precedence of causal variables. As O'Leary, Malone, and Tyree (1994) noted in their discussion of marital violence, it is difficult to draw inferences about causal factors unless they are measured well ahead of the targeted behavior problem. A

concurrent measurement strategy (in which hypothesized independent and dependent variables are measured concurrently) cannot be sufficient for causal inferences because the form (e.g., correlational, bidirectional causal) of the relationships cannot be established.

Statistical analysis of time-series data can be cumbersome in many clinical assessment contexts (however, see subsequent discussion of computer aids). However, time-series assessment is also the best strategy for tracking the time-course of behavior problems and goal attainment during therapy. Frequent monitoring increases the chance that failing treatment will be detected early (Mash & Hunsley, 1993) and that the clinician can make needed changes in the functional analysis and intervention programs.

Time-series assessment is congruent with an emphasis on professional accountability and empirically based judgments in clinical treatment. Time-course plots can be a useful means of documenting treatment effects and of providing feedback to clients about behavior change and possible causal relationships for their behavior problems. It can help the clinician identify naturally occurring causal mechanisms for a client's mood fluctuations, panic episodes, and other behavior problems. Time-series data was acquired with Mrs. A when she self-monitored daily her headaches and sleep problems.

Finally, time-series measurement is an essential element in interrupted time-series designs, such as the A-B-A-B, multiple baseline, or changing-criterion designs (Kazdin, 1997; Shapiro & Kratochwill, 1988). These designs can strengthen the internal validity of inferences about treatment effects and mechanisms.

4.07.4.4 Quantitative and Qualitative Approaches to Behavioral Assessment

The empirical elements of the behavioral assessment paradigm are partially responsible for the current emphasis on treatment outcome evaluation through frequently applied, minimally inferential, validated assessment instruments. This quantitative approach to clinical inference reflects and accentuates the growing importance of systematic evaluation by professionals delivering clinical services. The behavioral assessment paradigm provides a useful structure for the evaluation of clinical service delivery and intervention outcome. However, it is possible for clinicians and clinical researchers to overemphasize excessively molecular measures and excessive quantification. Quantification is an indispensable component of clinical

inference but an exclusive reliance on quantification can promote a focus on variables and effects with minimal practical importance. The emphasis on clinical significance (e.g., Jacobson & Truax, 1991) of treatment effects and functional relationships reflects the practical and clinical importance of effects, as well as their statistical significance.

Qualitative analyses (Haberman, 1978) can complement quantitative analyses. Behavioral assessors can generate clinically useful hypothesis by supplementing quantitative with qualitative analyses of clinical phenomena. Using time-sampling measurement strategies to code specific dyadic interaction sequences in a communication analogue between Mrs. and Mr. A provided data that helped identify dysfunctional verbal exchanges and provided the data base for judging the effects of communication training. However, it was also beneficial for the clinician to "passively" listen to the couple discuss their marital problems. Qualitative observation can be a rich source of ideas about the beliefs, attitudes, and behaviors that may contribute to marital distress.

Qualitative analyses can also promote the development of the behavioral assessment paradigm. We have only an elementary understanding of the causes of behavior disorders and of the best methods of treating them. An exclusive reliance on quantification can impair the evolution of any nascent assessment paradigm. Consequently, we must adopt a Steinbeckian attitude—generate and consider new ideas about functional relationships, presume that a new idea may be true, and then rigorously examine it.

Although I am advocating that qualitative methods can contribute to behavioral assessment, scientific methods remain the core element in the behavioral assessment paradigm. The stagnant nature of many psychological construct systems can be attributed to their focus on a rigidly invoked core set of assumptions about the nature of behavior and treatment, rather than on a core set of beliefs about the best way to learn about behavior and treatment. The behavioral assessment paradigm will continue to evolve to the degree that it emphasizes scientific methods for studying behavior, rather than a set of *a priori* beliefs about the nature of behavior disorders and their treatment.

4.07.5 BEHAVIORAL ASSESSMENT METHODS

There are many methods of behavioral assessment. Some, such as self-monitoring

and behavioral observation, are congruent with and influenced by the conceptual and methodological elements of the behavioral assessment paradigm outlined earlier in this chapter. Others, such as trait-focused self-report questionnaires, are less congruent with the behavioral assessment paradigm. Surveys of journal publications and of the assessment methods used by practicing behavior therapists show that it is difficult to reliably categorize an assessment instrument as “behavioral” or “nonbehavioral”—categories of behavioral and nonbehavioral assessment methods are becoming increasingly indistinct. For example, many cognitive assessment instruments used by behavior therapists are aggregated and trait-based. They provide an unconditional “score” of some cognitive construct such as “self-efficacy beliefs” or “locus of control” (see discussion of cognitive assessment in Linscott & DiGiuseppe, 1998). Other assessment instruments and methods used by behavior therapists include neuropsychological assessment, sociometric status evaluation, the Minnesota Multiphasic Personality Inventory and other trait-based personality tests, aggregated mood scales, historically focused interviews, and tests of academic achievement (see Hersen & Bellack, 1998).

There are several bases for the decreasing distinction between behavioral and nonbehavioral assessment methods. First, many variables currently targeted in behavioral assessment (e.g., subjective distress, beliefs, mood) were not the focus of behavior therapy several decades ago. As the response modes in causal models of behavior disorders and the targets of behavioral treatments expanded beyond motor and verbal behavior, the array and focus of assessment instruments used in behavioral assessment expanded correspondingly.

Second, behavioral assessors are less predisposed to immediately reject all traditional self-report assessment instruments. Some well-validated “personality” assessment instruments can provide useful information. However, care should be taken to avoid unwarranted inferences from personality assessment instruments and to insure that, if used, they are part of an assessment program that includes more precisely focused assessment methods (see discussions of behavioral and personality assessment in *Behavior Modification*, 1993, No. 1, and general problems in personality assessment in Heatherton & Weinberger, 1994).

Third, in the 1960s and 1970s many behavior analysts denounced trait concepts and emphasized situational sources of behavior variance. The current person $\times$ situation interactional model of behavior variance, discussed earlier,

suggests that clinical judgments can sometimes be aided by some trait measures, when used with situationally sensitive assessment instruments.

Fourth, the power of behavioral assessment methods often surpasses their clinical utility. Some behavioral assessment methods, such as behavioral observation, are costly and time-consuming, which decreases their usefulness in clinical assessment (however, as discussed in subsequent sections, many technological advances have enhanced the clinical utility of some behavioral assessment methods).

Finally, behavioral assessors are sometimes insufficiently educated in psychometric principles and in the degree to which assessment strategies match the conceptual and methodological aspects of the behavioral assessment paradigm. For example, behavioral assessors sometimes use an aggregated score from an assessment instrument that has multiple uncorrelated factors. At other times an aggregated score is used when there is important between-situation variance in the measured construct. Also, norm-referenced assessment instruments are sometimes applied when they are not psychometrically appropriate or useful with a particular client (norms may not be available for the client’s ethnic group, age, or economic status; Cone, 1996; Haynes & Wai’alae, 1995; Silva, 1993).

The following sections briefly present four categories of behavioral assessment methods: (i) behavioral observation, in the natural environment and in analogue situations; (ii) self-monitoring; (iii) self-report interviews and questionnaires; and (iv) psychophysiological assessment. The specific strategies, conceptual foundations, utility, psychometric properties, disadvantages, technical advancements, and contribution to clinical judgment of each category will be discussed. Coverage of these methods is necessarily limited and more extensive presentations of behavioral assessment methods and instruments can be found in books by Hersen and Bellack (1998), Mash and Terdal (1988), Ollendick and Hersen (1993), and Shapiro and Kratochwil (1988).

4.07.5.1 Behavioral Observation

Behavioral observation involves the time-series observation of molecular, precisely defined behaviors, and environmental events. Usually, an observation session is divided into smaller time samples and the occurrence of discrete events within each time sample is recorded by external observers (Foster & Cone, 1986; Foster, Bell-Dolan, & Burge, 1988; Mash & Hunsley, 1990; Tryon, 1998).

Two observation strategies, and variants of each, are discussed below: behavioral observation in the natural environment and behavioral observation in analogue environments.

4.07.5.1.1 Behavioral observation in the natural environment

Systematic behavioral observation of clients is congruent with most of the underlying assumptions of the behavioral assessment paradigm. Quantitative, minimally inferential data are derived on clients in their natural environment using external observers (nonparticipant observers). Behavior observation systems can be constructed for individual patients, some sources of measurement error can be examined through interrater reliability coefficients, and the acquired data can provide valuable information for the functional analysis and for treatment outcome evaluation. Observation in the natural environments has been used in the assessment of self-injurious, delusional, and hallucinatory behaviors in institutions; eating and drinking in restaurants, bars, and in the home; marital and family interactions in the home; student, teacher, and peer interactions in schools; pain and other health-related behaviors at home and in medical centers; community-related behaviors (e.g., driving, littering); and many other behaviors.

Typically, the client is observed for brief periods (e.g., one hour) in his or her natural environment (e.g., at home) several times in “sensitive” or “high-risk” situations—situations or times with an elevated probability that important behaviors or interactions will occur (e.g., at dinnertime when observing problematic family interactions; at mealtime when observing the social interactions of a psychiatric inpatient). Trained observers record the occurrence/nonoccurrence of specified and carefully defined behaviors within *time-sampling* periods (e.g., 15-second periods). Sequences of events (e.g., sequential interactions between a depressed client and family members) and the duration of events can also be recorded. Observers can also use momentary time sampling and record the behaviors that are occurring at predetermined time sampling points. An example of this latter sampling strategy would be a nurse recording the social interactions of a psychiatric inpatient at the beginning of every hour.

Observation in unrestricted environments (e.g., in a client’s home) can be problematic because clients are sometimes out of sight of the observers and sometimes engage in behaviors that are incompatible with the purposes of the observation (e.g., talking on the phone; watch-

ing TV). Consequently, the behavior of the individual to be observed is often constrained. For example, a marital couple being observed at home might be requested to remain within sight of the observer, to avoid long phone conversations, and keep the TV off. Such constraints compromise the generalizability of the obtained data but increase the efficiency of the observation process.

The temporal parameters (the length, frequency, spacing, and sampling intervals) of an observation session are influenced by the characteristics of the targeted behaviors. For example, higher rate behaviors require shorter time sampling intervals. Highly variable behaviors required more observation sessions. Suen and Ary (1989) discuss temporal parameters of behavioral observation in more detail.

Behavioral observation in the natural environment has many applications. It is a powerful method of treatment outcome evaluation because it minimizes many sources of measurement error associated with self-report and is sensitive to behavior change. It has also been used as a basic research tool in the behavioral and social sciences and to gather data for functional analyses.

The clinical utility of behavioral observation in the natural environment is limited in several ways. It is not cost-efficient for the assessment of very low frequency behaviors, such as stealing, seizures, and some aggressive behaviors. Also, “internalized,” covert behavior problems such as mood disorders and anxiety are less amenable to external observation (although some verbal and motoric components of these disorders can be observed). Socially sensitive or high-valence behaviors, such as sexual dysfunctions, paraphiliac behaviors, substance use, and marital violence may not be emitted in the presence of observers. In most outpatient clinical settings, observation in the natural environment with external observers is prohibitively expensive and time-consuming.

Behavioral observation is an expensive assessment method but technological advances have enhanced its clinical utility. Audio and video tape recorders and other instrumentation can facilitate the acquisition of observation data on clients in the natural environment without having to send observers (Tryon, 1991). Observers can also use hand-held computers to record and analyze observation data in real-time (Tryon, 1998).

Because many behaviors can be observed, behavior sampling is an indispensable component of behavioral observation. Typically, observers use a standardized behavior coding system (e.g., Marital Interaction Coding System; Weiss & Summers, 1983) that contains

preselected behavior codes. Behaviors selected for inclusion in behavioral observation include: (i) client problem behaviors (e.g., social initiation behaviors by a depressed psychiatric inpatient), (ii) causal variables for the client's behavior problems and goals (e.g., compliments and insults emitted during distressed marital interaction), (iii) behaviors correlated with client problem behaviors (e.g., members of a functional response class such as verbal and physical aggressive behaviors), (iv) salient, important, high-risk behaviors (e.g., suicide talk), (v) client goals and positive alternatives to undesirable behaviors (e.g., positive social interactions by a delusional or socially isolated client), (vi) anticipated immediate, intermediate, and final outcomes of treatment, and (vii) possible positive and negative side- and generalized effects of treatment.

Although observers sometimes focus on only one individual, it is more common in behavioral observation to monitor interpersonal interactions. To this end, observers can record sequences of interactions between two or more individuals (see discussions in Bakeman & Gottman, 1986; Moran et al., 1992). When the goal of observation is to draw inferences about a group of persons, subject sampling can be used. A few persons may be selected for observation from a larger group. For example, several patients may be selected for observation if the goal is to document the effects of a new token program on patients in a psychiatric unit.

Behavioral observation is often considered the "gold standard" for assessment. However, there are several sources of error which can detract from the accuracy and validity of obtained data and of the inferences drawn from them. Sources of error in behavioral observation include: (i) the degree to which observers are trained, (ii) the composition and rotation of observer teams, (iii) observer bias and drift, (iv) the behaviors selected for observation, (v) the specificity of code definitions, (vi) the methods of evaluating interobserver agreement, (vii) the match between time samples and the dynamic characteristics of the observed behaviors, (viii) variance in the situations or time of day in which observation occurs (Alessi, 1988; Hartmann, 1982; Suen & Ary, 1989; Tryon, 1996).

Several types of data can be derived from behavioral observation. Often, assessors are interested in the rate of targeted events. This is usually calculated as the percent of sampling intervals in which a behavior occurs. Sequential analyses and conditional probabilities are often more important for developing functional analyses. For example, observation of family interaction in the home can provide data on

negative reciprocity—the relative probability that one family member will respond negatively following a negative (in comparison to a nonnegative) response by another family member. Particularly with computer-aided observation, data on the duration of behaviors can also be acquired. Some observation systems use ratings by observers, rather than event recordings, although these are more frequently conducted outside formal observation session (Segal & Fal, 1998; Spector, 1992). According to McReynolds (1986), the first rating scale was published by Thomasius in 1692. Four basic characterological dimensions were rated—sensuousness, acquisitiveness, social ambition, and rational love.

"Reactivity" refers to the degree to which assessment modifies the targets of assessment. Reactivity is a potential source of inferential error in all assessment procedures, but particularly in behavioral observation (Foster et al., 1988; Haynes & Horn, 1982). The behavior of clients, psychiatric staff members, spouses, and parents may change as a function of whether observers are present or absent. Therefore, reactivity threatens the external validity or situational and temporal generalizability of the acquired data. In the cases of some highly socially sensitive behaviors (e.g., sexual or antisocial behaviors), the magnitude of reactivity may be sufficient to preclude observation in the natural environment.

Participant observation is an alternative to the use of nonparticipant observers. Participant observation is behavioral observation, as described above, using observers who are normally part of the client's natural environment. (In ethnography and cultural anthropology, the term "participant observation" usually refers to qualitative observation by external observers who immerse themselves in the social system they are observing.) Examples of participant observation include: (i) parents observing the play behavior of their children, (ii) nurses observing the delusional speech of psychiatric inpatients, and (iv) a spouse observing a client's depressive or seizure behaviors. Participant observers often use time and behavior sampling procedures similar to those used by nonparticipant observers. However, participant observers are usually less well trained and apply less complex observation systems (e.g., fewer codes, use of momentary time sampling). For example, a staff member on a psychiatric unit might monitor the frequency of social initiations by a client only during short mealtime periods or only briefly throughout the day.

The primary advantages of participant observation are its cost-efficiency and applicability. Participant observation can be an

inexpensive method of gathering data on clients in their natural environment. It may be particularly advantageous for gathering data on low frequency behaviors and on behaviors that are highly reactive to the presence of external observers.

There are several potential sources of error in participant observation. First, it is susceptible to most of the sources of error mentioned for nonparticipant observation (e.g., behavior definitions, time sampling errors). Additionally, participant observation may be particularly sensitive to observer biases, selective attention by observers, and recent interactions with the target of observation. The observer is likely to be less well trained and often is not a neutral figure in the social context of the client.

Participant observation may be associated with reactive effects. Sometimes, the reactive effects would be expected to be less for participant than for nonparticipant observation. One determining factor in reactivity is the degree to which the assessment process modifies the natural environment of the client. Because participant observation involves less change in the natural environment of the client, it may be less reactive. However, the reactive effects of participant observation may be strengthened, depending on the method of recording, the behaviors recorded, and the relationship between the observer and client. In some situations participant observation might be expected to alter the monitored behavior to an important degree or to adversely affect the relationship between the observer and target (e.g., an individual monitoring the sexual or eating behavior of a spouse).

Critical event sampling is another infrequently used but clinically useful and cost-efficient method of acquiring natural environment observation data. Critical event sampling involves video or audio tape recording of important interactions in the client's natural environment (Jacob, Tennenbaum, Bargiel, & Seilhamer, 1995; Tryon, 1998). The critical interactions are later qualitatively or quantitatively analyzed. For example, a tape recorder could be self-actuated by a distressed family during mealtime; a marital couple could record their verbal alterations at home; a socially anxious individual could record conversations while on a date.

4.07.5.1.2 Analogue observation

Analogue observation involves the systematic behavioral observation of clients in carefully structured environments. The assessment environment is arranged to increase the probability that clinically important behaviors and

functional relationships can be observed. It is a powerful, clinically useful, and idiographic behavioral assessment method. Many elements of the analogue assessment setting are similar to those of the client's natural environment. However, the participants, social and physical stimuli, and instructions to the client may differ from those of the client's natural environment.

The behavior of clients in analogue assessment is presumed to correlate with their behavior in the natural environment. For example, a distressed marital couple might discuss a problem in their relationship while in a clinic and being observed from a one-way mirror. It is presumed that the problem-solving strategies they use will be similar to those they use outside the clinic.

Analogue observation has many applications. The role play is often used in the behavioral assessment of social skills. A psychiatric patient or socially anxious adult might be observed in a clinic waiting room while attempting to initiate and maintain a conversation with a confederate. A client might be placed in a simulated social situation (e.g., simulated restaurant) and asked to respond to social stimuli provided by a confederate. The Behaviour Avoidance Test (BAT; e.g., Beck & Zebb, 1994) is another form of analogue observation. In a BAT, a client is prompted to approach an object that is feared or avoided.

Analogue methods have been used in the assessment of many clinical phenomena, such as pain (Edens & Gil, 1995), articulated thoughts (Davison, Navarre, & Vogel, 1995), and social anxiety and phobia (Newman, Hofmann, Trabert, Roth, & Taylor, 1994). Other applications include the study of self-injurious behaviors, dental anxiety, stuttering, heterosexual anxiety, alcohol ingestion, panic episodes, cigarette refusal skills, parent-child interaction, speech anxiety, animal phobias, test anxiety, and eating patterns.

Data can be acquired on multiple response modes in analogue observation. For example, during exposure to an anxiety provoking social exchange, clients can report their level of anxiety and discomfort, electrophysiological measures can be taken, observers can record the behavior of the client, and clients can report their thoughts.

Analogue observation is a cost-efficient and multimodal method of assessment and can be a useful supplement to retrospective interview and questionnaire methods. It provides a means of directly observing the client in sensitive situations and of obtaining *in vivo* client reports to supplement the client's retrospective report of thoughts, emotions, and behavior. In comparison to observation in the natural

environment, it is particularly useful for observing some important but low-rate events (e.g., marital arguments).

When used in conjunction with systematic manipulation of hypothesized controlling variables analogue observation can be exceptionally useful for identifying causal relationships and for constructing a functional analysis of behavior problems. For example, social attention, tangible rewards, and task demands can be systematically presented and withdrawn before and after the self-injurious behavior of developmentally disabled individuals (e.g., Iwata et al., 1994). Systematic manipulation of hypothesized controlling factors during analogue observation has also been used to identify the cognitive sequelae of hyperventilation during anxiety episodes (Schmidt & Telch, 1994), the most effective reinforcers to use in therapy (Timberlake & Farmer-Dougan, 1991), and the factors associated with food refusal (Munk & Repp, 1994).

Analogue observation is associated with several unique sources of variance, measurement error, and inferential error (e.g., Hughes & Haynes, 1978; Kern, 1991; Torgrud & Holborn, 1992). First, because the physical environment and social stimuli are more carefully controlled in analogue observation but may differ from those in the client's naturalistic environment, external validity may be reduced concomitantly with an increase in behavioral stability. The behavior of clients and the data acquired during analogue observation can covary with: (i) the physical characteristics of the assessment environment; (ii) the instructions to participants; (iii) observer and sampling errors, such as those outlined in naturalistic observation, time and behavior sampling; and (iv) the content validity of the assessment environment (i.e., the degree to which the stimuli are relevant to the construct being measured by the analogue assessment situation).

The primary disadvantage to analogue observation is its representational nature: the data acquired in analogue assessment are only presumed to correlate with data that would be acquired in the natural situations the analogue assessment is designed to represent. It is an indirect measure of the individual's behavior in the natural environment. The results of many studies have supported the discriminant and criterion-related validity of analogue observation; the results of other validation studies have suggested more cautious conclusions. Given the presumption that many behaviors are sensitive to situational sources of variance, clients can be expected to behave differently in analogue and natural environments. That is, the behavior of clients in

analogue settings may not accurately reflect or match their behavior in the natural environment. Nevertheless, analogue assessment should be expected to be valid in other ways. For example, socially anxious and nonanxious clients should behave differently during analogue observation even if their behaviors do not match their behaviors in the natural environment. The validity and clinical utility of analogue observation should be considered dependent variables. They are likely to vary across the purposes of the assessment, subjects, target behaviors, settings, and observation methods.

4.07.5.2 Self-monitoring

In self-monitoring, clients systematically record their behavior and sometimes relevant environmental events (Bornstein, Hamilton, & Bornstein, 1986; Gardner & Cole, 1988; Shapiro, 1984). The events to be recorded by the client are first identified and specified by the client and assessor during an interview. A recording form is developed or selected and the client monitors the designated events, sometimes at designated times or in specified situations. To reduce errors associated with retrospective reports, recording usually occurs immediately before or after the monitored event.

One particularly innovative development is self-monitoring via hand-held computer and computer-assisted data acquisition and analysis (Agras, Taylor, Feldman, Losch, & Burnett, 1990; Shiffman, 1993). Hand-held computers allow the collection of real-time data and simplify the analysis and presentation of self-monitoring data. Computerization should significantly increase the clinical utility of self-monitoring.

Time-sampling is sometimes used with self-monitoring, depending on the characteristics of the behavior. Clients can easily record every occurrence of very low-rate behaviors such as seizures or migraine headaches. However, with high-rate or continuous behaviors, such negative thoughts, blood pressure, and mood, clients recordings may be restricted to specified time periods or situations.

Many clinically important behaviors have been the targets of self-monitoring. These include ingestive behaviors (e.g., eating, caffeine intake, alcohol and drug intake, smoking), specific thoughts (e.g., self-criticisms); physiological phenomena and medical problems (e.g., bruxism, blood pressure, nausea associated with chemotherapy, Raynaud's symptoms, arthritic and other chronic pain, heart rate, seizures); and

a variety of other phenomena such as self-injurious behaviors, electricity use, startle responses, sexual behavior, self-care behaviors, exercise, panic and asthma episodes, social anxiety, mood, marital interactions, study time, sleeping patterns, and nightmares.

Many response modes are amenable to measurement with self-monitoring. Clients can monitor overt motor behavior, verbal behavior, subjective distress and mood, emotional responses, occurrence of environmental events associated with their behavior, physiological responses, thoughts, and the qualitative characteristics of their behavior problems (e.g., location of headaches, specific negative thoughts, multimodal aspects of panic episodes). Self-monitoring can also be used to track multiple response parameters: response durations, magnitudes, and frequencies. Of particular relevance for the functional analysis, the client can concurrently monitor behaviors, antecedent events, and consequent events to help identify functional relationships. Johnson, Schlundt, Barclay, Carr-Nangle, and Engler (1995), for example, had eating disordered clients monitor the occurrence of binge eating episodes and the social situations in which they occurred in order to calculate conditional probabilities for binge eating.

Self-monitoring is a clinically useful assessment method. Self-monitoring can be tailored for individual clients and used with a range of behavior problems. It is an efficient and inexpensive assessment method for gathering data on functional relationships in the natural environment and is another important supplement to retrospective self-report. It is suitable for time-series assessment and for the derivation of quantitative indices of multiple response modes. Self-monitoring is applicable with many populations—adult outpatients, children, inpatients, parents and teachers, and developmentally disabled individuals. Events that, because of their frequency or reactive effects, are not amenable to observation by participant and nonparticipant observers may be more amenable to assessment with self-monitoring.

Although many validation studies on self-monitoring have been supportive (see reviews in Bornstein et al., 1986; Gardner & Cole, 1988; Shapiro, 1984) there are several threats to the validity of this assessment method. Two important sources of error in self-monitoring are clients' recording errors and biases. The resultant data can reflect the client's abilities to track and record behaviors, client expectancies, selective attention, missed recording periods, the social valence and importance of the target behaviors, fabrication, and the contingencies associated with the acquired data. Data can also

be affected by how well the client was trained in self-monitoring procedures, the demands of the self-monitoring task, the degree to which target events have been clearly specified, reactions from family and friends to the client's monitoring, and the frequency and duration of the targeted behaviors. Clinical researchers have also reported difficulty securing cooperation from clients to self-monitor for extended periods of time. One particularly powerful source of inferential error is reactivity (Bornstein et al., 1986). The reactive effects of self-monitoring are frequently so great that self-monitoring is sometimes used as a method of treatment.

4.07.5.3 Psychophysiological Assessment

Psychophysiological measurement is an increasingly important method in behavioral assessment (Haynes, Falkin, & Sexton-Radek, 1989). The increased emphasis on psychophysiological assessment is due, in part, to a growing recognition of the importance of physiological components of behavior problems, such as depression, anxiety, and many psychotic behavior problems. Also, behavior therapists are increasingly involved in the assessment and treatment of disorders that have traditionally been the focus of medical interventions—cancer, chronic pain, diabetes, cardiovascular disorders. A third reason for the importance of psychophysiological assessment is that many behavioral intervention procedures, such as relaxation training and desensitization, focus partly on the modification of physiological processes. Advances in ambulatory monitoring, computerization, and other technologies have increased the clinical utility of psychophysiological measurement. Finally, psychophysiological measurement can easily be combined with other behavioral assessment methods, such as self-monitoring and analogue observation.

The recognition of the importance of the physiological response mode in behavior problems mandates the inclusion of electrophysiological and other psychophysiological measurement methods. Electromyographic, electrocardiovascular, electroencephalographic, and electrodermal measures are particularly applicable to behavioral assessment with adults. A range of behavior problems (e.g., panic disorders, PTSD, schizophrenic behaviors, obsessive-compulsive behaviors, worry, depression, substance abuse, disorders of initiating and maintaining sleep) have important physiological components. The low magnitude of covariance between physiological and other response modes,

noted earlier in this chapter, suggests that they may be a function of different causal variables and respond differently to the same treatment.

Psychophysiological measurement is a complex, powerful, and clinically useful assessment method in many assessment contexts and for many clients. It is amenable to idiographic assessment, can be applied in a time-series format, and generates quantitative indices. The validity of the obtained measures can be affected by electrode placement, site resistance, movement, instructional variables, time-sampling parameters, data reduction and analysis, equipment intrusiveness, and equipment failures. Books by Andreassi (1995) and Cacioppo and Tassinary (1990) cover instrumentation, measurement methods, technological innovations, clinical applications, and sources of measurement error.

4.07.5.4 Self-report Methods in Behavioral Assessment

Many interview formats and hundreds of self-report questionnaires have been adopted by behavioral assessors from other assessment paradigms. A comprehensive presentation of these methods is not possible within the confines of this chapter. Here, I will emphasize how behavioral and traditional self-report methods differ in format and content. The differences reflect the contrasting assumptions of behavioral and nonbehavioral assessment paradigms.

More extensive discussions of self-report questionnaire and interview methods, and applicable psychometric principles are provided by Anastasi (1988), Jensen and Haynes (1986), Nunnally and Bernstein (1994), Sarwer and Sayers (1998), and Turkat (1986).

Behavioral assessors, particularly those affiliated with an applied behavior analysis paradigm, have traditionally viewed self-report questionnaires and interviews with skepticism. Objections have focused on the content and misuses of these methods. Many questionnaires solicit retrospective reports, stress situationally insensitive aggregated indices of traits, focus on molar level constructs that lack consensual validity, and are unsuited for idiographic assessment. Biased recall, demand factors, item interpretation errors, and memory lapses, further challenged the utility of self-report questionnaires. Data from interviews have been subject to the same sources of error with additional error variance associated with the behavior and characteristics of the interviewer. Despite these constraints, interviews and questionnaires are the most frequently used methods used by behavior therapists (e.g., Piotrowski & Zalewski, 1993).

The interview is an indispensable part of behavioral assessment and treatment and undoubtedly is the most frequently used assessment method. All behavioral interventions require prior verbal interaction with the client or significant individuals (e.g., staff) and the structure and content of that interview can have an important impact on subsequent assessment and treatment activities.

As illustrated with Mrs. A, an assessment interview can be used for multiple purposes. First, it can help identify and rank order the client's behavior problems and goals. It can also be a source of information on the client's reciprocal interactions with other people, and, consequently, provides important data for the functional analysis. Interviews are the main vehicles for informed consent for assessment and therapy and can help establish a positive relationship between the behavior assessor and client. Additionally, interviews are used to select clients for therapy, to determine overall assessment strategies, to gather historical information, and to develop preliminary hypotheses about functional relationships relevant to the client's behavior problems and goals.

The behavioral assessment interview differs from nonbehavioral interviews in content and format. First, the behavioral interview is often more quantitatively oriented and structured (although most behavioral interviews involve unstructured, nondirective, and client-centered phases). The focus of the behavioral interview reflects assumptions of the behavioral assessment paradigm about behavior problems and causal variables and emphasizes current rather than historical behaviors and determinants. Behavioral interviewers are more likely to query about situational sources of behavioral variance and to seek specification of molecular behaviors and events.

A systems perspective also guides the behavioral assessment interview. The behavioral interviewer queries about the client's extended social network and the social and work environment of care-givers (e.g., the incentives at a psychiatric institution that encourage cooperation by staff members). The interviewer also evaluates the effects that treatment may have on the client's social system—will treatment effect family or work interactions?

Some of the concerns with the interview as a source of assessment information reside with its traditionally unstructured applications. Under unstructured conditions, data derived from the interview may covary significantly with the behavior and biases of the interviewer. However, structured interviews and technological advances in interview methods promise to reduce such sources of error (Hersen & Turner,

1994; Sarwer & Sayers, 1998). Computerization, to guide the interviewer and as an interactive system with clients, promises to reduce some sources of error in the interview process. Computerization can also increase the efficiency of the interview and assist in the summarization and integration of interview-derived data.

Other structured interview aids, such as the Timeline Followback (Sobell, Toneatto, & Sobell, 1994) may also increase the accuracy of the data derived in interviews. In Timeline Followback, memory aids are used to enhance accuracy of retrospective recall of substance use. A calendar is used as a visual aid, with the client noting key dates, long periods in which they abstained or were continuously drunk, and other discreet events associated with substance use.

Some interviews are oriented to the information required for a functional analysis. For example, the Motivation Assessment Scale is used with care-givers to ascertain the factors that may be maintaining or triggering self-injurious behavior in developmentally disabled persons (Durand & Crimmins, 1988).

Questionnaires, including rating scales, self-report questionnaires, and problem inventories, are also frequently used in behavioral assessment; they have probably been frequently used with all adult behavior disorders. Many questionnaires used by behavioral assessors are identical to those used in traditional nonbehavioral psychological assessment. As noted earlier, questionnaires are often adopted by behavioral assessors without sufficient thought to their underlying assumptions about behavior and the causes of behavior problems, content validity, psychometric properties, and incremental clinical utility. They are often trait-focused, insensitive to the conditional nature of the targeted behavior, and provide aggregated indices of a multifaceted behavioral construct (Haynes & Uchigakiuchi, 1993). Questionnaires are sometimes helpful for initial screening or as a nonspecific index of program outcome but are not useful for a functional analysis or for precise evaluation of treatment effects. The integration of personality and assessment is addressed further in a subsequent section of this chapter.

Some questionnaires are more congruent with the behavioral assessment paradigm. These usually target a narrower range of adult behavior problems or events, such as panic and anxiety symptoms, outcome expectancies for alcohol, recent life stressors, and tactics for resolving interpersonal conflicts. Most behaviorally oriented questionnaires focus on specific and lower-level behaviors and events and query about situational factors. However, the developers of behaviorally oriented questionnaires

sometimes rely on the face validity of questionnaires and do not follow standard psychometric principles of questionnaire development (see special issue on "Methodological issues in psychological assessment research" in *Psychological Assessment*, 1995, Vol. 7). Deficiencies in the development and validation of any assessment instrument reduce confidence in the inferences that can be drawn from resulting scores.

Questionnaires, given appropriate construction and validation, can be an efficient and useful source of behavioral assessment data. Most are inexpensive, quick to administer and score, and are well received by clients. Computer administration and scoring can increase their efficiency and remove several sources of error (Honaker & Fowler, 1990). They can be designed to yield data on functional relationships of variables at a clinically useful level of specificity.

4.07.5.5 Psychometric Foundations of Behavioral Assessment

The application of psychometric principles to behavioral assessment instruments has been discussed in many books and articles (e.g., Cone, 1988; 1996; Foster & Cone, 1995; Haynes & Wai'alaie, 1995; Silva, 1993; see also "Methodological issues in psychological assessment research," *Psychological Assessment*, September, 1995). Psychometric principles were originally applied to tests of academic achievement, intelligence, and abilities. Because many of the principles were based on estimating measurement error with presumably stable and molar-level phenomena, the relevance of psychometric principles to behavioral assessment has been questioned. However, "psychometry" is best viewed as a general validation process that is applicable to any method or instrument of psychological assessment.

The ultimate interest of psychometry is the construct validity of an assessment instrument or, more precisely, the construct validity of the data and inferences derived from an assessment instrument. Construct validity is comprised of the multiple lines of evidence and rationales supporting the trustworthiness of assessment instrument data interpretation (Messick, 1993). Indices of construct validity are also conditional—an index of validity does not reside unconditionally with the instrument (Silverman & Kurtines, 1998, discuss contextual issues in assessment). Elements of construct validation are differentially applicable, depending on the method, target, and purpose of assessment.

The validity of data derived from an assessment instrument establishes the upper limit of confidence in the clinical judgments to which the instrument contributes. Consequently, the validity of every element of the functional analysis is contingent on the validity of the assessment instruments used to collect contributing data. The validity of other clinical judgments (e.g., risk factors for relapse, and the degree of treatment effectiveness) similarly depends on the validity of assessment data.

The applicability of psychometric principles (e.g., internal consistency, temporal stability, content validity, criterion-related validity) to behavioral assessment instruments varies with their methods, targets, and applications. The data obtained in behavioral assessment differ in the degree to which they are presumed to measure lower-level less-inferential variables (e.g., number of interruptions in a conversation, hitting) or higher-level more inferential variables (e.g., positive communication strategies, aggression). With lower-level variables, psychometric indices such as internal consistency and factor structure are not useful indications of validity of the obtained data. Interobserver agreement and content validity may be more useful indices.

The validity of data from an assessment instrument depends on how it will be used—on the clinical judgments that it affects. For example, accurate data may be obtained from analogue observation of clients social interactions. That is, there may be perfect agreement among multiple observers about the client's rate of eye contact, questions, and reflections. However, those rates may demonstrate low levels of covariance (i.e., low criterion-referenced validity) with the same behaviors measured in natural settings. The relative importance of accuracy and other forms of validity varies with the purpose of the assessment (see Cone, 1998). If the analogue data is used to evaluate the effectiveness of a social skills training program, accuracy is an important consideration. If the data is to be used to evaluate generalization of treatment effects, accuracy is necessary but not sufficient.

The interpretation of temporal and situational stability coefficients is complicated in behavioral assessment by the conditional and unstable nature of some of the targeted phenomena (e.g., appetitive disorders, social behaviors, mood, expectancies). Indices of instability (e.g., low test-retest correlations) can reflect variability due to true change across time in the variable (e.g., change in the social behavior of an observed client) as well as measurement error (e.g., poorly defined behavior codes, observer error). Consequently,

temporal stability coefficients, by themselves are weak indices of validity. A multimethod/multi-instrument assessment strategy, by providing indices of covariance among measures of the same targeted phenomena, however, can help separate true from error variance. Additionally, a low magnitude of temporal stability in a time-series measurement strategy has implications for the number of samples necessary to estimate or capture the time course of the measured phenomena—unstable phenomena require more sampling periods than do stable phenomena.

Behavioral assessment often involves multiple methods of assessment, focused on multiple modes and parameters. As noted earlier in this chapter, sources of measurement error and determinants can vary across methods, modes, and parameters. A multimethod approach to assessment can strengthen confidence in subsequent clinical judgments. However, estimates of covariance are often used as indices of validity and can be attenuated in comparison to monomethod or monomode assessment strategies (see discussion of psychometric indices of multiple methods in Rychtarik & McGillicuddy, 1996).

The individualized nature of behavioral assessment enhances the importance of some construct validity elements. For example, accuracy, content validity, and interobserver agreement are important considerations in behavioral observation coding systems. Idiographic assessment reduces the importance of construct validity elements such as nomothetically based discriminant and convergent validity.

4.07.6 BEHAVIORAL AND PERSONALITY ASSESSMENT

As noted earlier in this chapter, behavioral assessors often use traditional personality questionnaires and several possible reasons for this integration were given. The positive cost-efficiency of personality trait measures is one factor. One of the more empirically based rationales for integration is a person $\times$ situation interactional model for assessment: if we want to predict a person's behavior, it helps to know something about the relatively stable aspects of the person and something about the situations that promote instability, at least sometimes (McFall & McDonel, 1986). Personality questionnaires are often used in initial screening, followed by more specifically focused, molecular, and less inferential assessment instruments. Noted in this section are several additional issues concerning the integration

of personality and behavioral assessment. These issues were discussed in Haynes and Uchigakiuchi (1993) and in other articles in a special section of *Behavior Modification*, 1993, 17(1).

There are several complications associated with adopting the situation $\times$ person interaction model and with the use of personality assessment instruments. Given that there are hundreds of traits measurable by extant instruments, it is difficult to determine which traits to measure, and how best to measure them. Also, despite a growing literature on situational factors in behavior disorders, we still do not know which aspects of situations can be most important in controlling behavioral variance for a particular client (e.g., Kazdin, 1979). Nor do we know under which conditions a person-situation interaction model, as opposed to a situational or trait model, will assume the greatest predictive efficacy.

Several additional issues regarding the trait $\times$ situation interactional model of behavior and the utility of personality assessment strategies for behavioral assessment have already been discussed in this chapter and in many previously published articles. First, personality traits vary in theoretical connotations and the theoretical connotations of a trait measure influence its utility for behavioral assessment. Many constructs measured by personality assessment instruments have psychodynamic and intrinsically causal connotations. Some, such as “emotional instability,” “hardiness,” and “passive-aggressive,” refer to an internal state that is presumed to control observed behavior—these traits invoke causal models that are inconsistent with aspects of the behavioral assessment paradigm. In these cases “psychological processes” are inferred from cross-situational consistencies in behavior. In a circular fashion, the processes become explanations for the behaviors that are their indicators. The processes cannot be independently validated, are difficult to measure, and the inferential process can inhibit a scientific investigation of these behaviors.

Personality questionnaires invariably invoke molar-level traits whose interpretation require normative comparison. Consequently, trait measures are less amenable to idiographic assessment of lower-level variables. Clinical inferences about a person on a trait dimension are derived by comparing the person’s aggregated trait score to the trait scores of a large sample of persons. Such comparative inferences can be helpful but can also be in error if there are important differences between the person and the comparison group, such as on dimensions of gender, ethnicity, and age.

Most behavioral assessors would acknowledge that molar self-report measures can contribute to clinical inferences, when used within a multimethod assessment program and care is taken to address many sources of measurement and inferential error noted above. However, there are several other complications associated with personality assessment: (i) many traits measured by personality assessment instruments are poorly defined and faddish; (ii) molar variables are less likely than molecular variables to reflect the dynamic nature of behavior—they are momentary snap-shots of unstable phenomena; (iii) personality trait measures may be more useful for initial screening than for the construction of a detailed functional analysis and treatment planning; (iv) personality traits can also be conditional: their probability and magnitude can vary across situations; (v) inferences about a client’s status on a trait dimension varies across assessment instruments; and (vi) because of their aggregated nature, many response permutations can contribute to a particular score on a trait dimension.

In sum, the integration of person-situation interactional models and personality assessment in the behavioral assessment paradigm can benefit clinical judgments. However, this integration has sometimes occurred too readily, without the thoughtful and scholarly reflection characteristics of the behavioral assessment paradigm.

4.07.7 SUMMARY

Behavioral assessment is a dynamic and powerful assessment paradigm designed to enhance the validity of clinical judgments. One of the most important and complex clinical judgments in behavioral assessment functional analysis—a synthesis of the clinicians hypotheses about the functional relationships relevant to a clients behavior problems. The functional analysis is a central component in the design of behavior therapy programs. The behavioral assessment paradigm suggests that errors in clinical judgments can be reduced to the degree that the judgments are based on multiple assessment methods and sources of information, validated assessment instruments, time-series measurement strategies, data on multiple response modes and parameters, minimally inferential variables, and the assessment of behavior-environment interactions. The Clinical Pathogenesis Map and Functional Analytic Causal Model were introduced as ways of graphically depicting and systematizing the functional analysis.

The methods of behavioral assessment and clinical case conceptualizations are influenced

by several interrelated assumptions about the causes of behavior problems. The behavioral assessment paradigm emphasizes multiple causality; multiple causal paths; individual differences in causal variables and paths; environmental causality and reciprocal determinism; contemporaneous causal variables; the dynamic nature of causal relationships; the operation of moderating and mediating variables; interactive and additive causality; situations, setting events, and systems factors as causal variables; and dynamical causal relationships.

The methods of behavioral assessment and clinical case conceptualizations are also affected by assumptions about the characteristics of behavior problems. These include an emphasis on the multimodal and multiparameter characteristics of behavior problems, differences among clients in the importance of behavior problem modes and parameters, the complex interrelationships among a client's multiple behavior problems, and the conditional and dynamic natures of behavior problems.

Three of many methodological foundations of behavioral assessment were discussed: the emphasis on empirical hypothesis-testing, the idiographic emphasis, and the use of time-series assessment strategies.

The decreasing distinctiveness of behavioral and nonbehavioral assessment, and reasons for this change, were discussed. Four categories of behavioral assessment methods were presented (i) behavioral observation, (ii) self-monitoring, (iii) self-report methods, and (iv) psychophysiological assessment. The specific strategies, conceptual foundations, clinical utility, psychometric properties, disadvantages, technical advancements, and contribution to clinical judgment of each category were presented.

The application of psychometric principles to behavioral assessment was discussed. The applicability of specific principles varies across methods, targets, and applications.

Several issues relating to the integration of behavioral and personality assessment were presented. These included poor definitions for some traits, the molar nature of personality assessment variables, insensitivity to dynamic aspects of behavior, reduced utility for functional analysis and treatment planning, the conditional nature of personality traits, differences among personality assessment instruments, and the aggregated nature of trait measures.

4.07.8 REFERENCES

- Agras, W. S., Taylor, C. B., Feldman, D. E., Losch, M., & Burnett, K. F. (1990). Developing computer-assisted therapy for the treatment of obesity. *Behavior Therapy*, 21, 99–109.
- Alessi, G. (1988). Direct observation methods for emotional/behavior problems. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 14–75). New York: Guilford Press.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Andreassi, J. L. (1995). *Psychophysiology: Human behavior and physiological response* (3rd ed.). Hillsdale, NJ: Erlbaum.
- Asterita, M. F. (1985). *The physiology of stress*. New York: Human Sciences Press.
- Bakeman, R., & Gottman, J. M. (1986). *Observing interaction: An introduction to sequential analysis*. New York: Cambridge University Press.
- Bandura, A. (1969). *Principles of behavior modification*. New York: Holt, Rinehart and Winston.
- Bandura, A. (1981). In search of pure unidirectional determinants. *Behavior Therapy*, 12, 315–328.
- Barlow, D. H., & Cerny, J. A. (1988). *Psychological treatment of panic*. New York: Guilford Press.
- Barnett, P. A., & Gotlib, I. H. (1988). Psychosocial functioning and depression: Distinguishing among antecedents, concomitants, and consequences. *Psychological Bulletin*, 104, 97–126.
- Barrios, B. A. (1988). On the changing nature of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (pp. 3–41). New York: Pergamon.
- Beach, S., Sandeen, E., & O'Leary, K. D. (1990). *Depression in marriage*. New York: Guilford Press.
- Beck, J. G., & Zebb, B. J. (1994). Behavioral assessment and treatment of panic disorder: Current status, future directions. *Behavior Therapy*, 25, 581–612.
- Bellack, A. S., & Hersen, M. (1988). *Behavioral assessment: A practical handbook*. New York: Pergamon.
- Bornstein, P. H., Bornstein, M. T., & Dawson, D. (1984). Integrated assessment and treatment. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 223–243). New York: Pergamon.
- Bornstein, P. H., Hamilton, S. B., & Bornstein, M. T. (1986). Self-monitoring procedures. In A. R. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 176–222). New York: Wiley.
- Brown, T. A., DiNardo, P. A., & Barlow, D. H. (1994). *Anxiety disorders interview schedule for DSM-IV (ADIS-IV)*. Albany, NY: Graywind Publications.
- Cacioppo, J. T., & Tassinari, L. G. (1990). *Principles and psychophysiology: Physical, social, and inferential elements*. New York: Cambridge University Press.
- Chadwick, P. D. J., Lowe, C. F., Horne, P. J., & Higson, P. J. (1994). Modifying delusions: The role of empirical testing. *Behavior Therapy*, 25, 35–49.
- Ciminero, A. R., Calhoun, K. S., & Adams, H. E. (1986). *Handbook of behavioral assessment*. New York: Wiley.
- Collins, L. M., & Horn, J. L. (Eds.) (1991). *Best methods for the analysis of change*. Washington, DC: American Psychological Association.
- Cone, J. D. (1979). Confounded comparisons in triple response mode assessment research. *Behavioral Assessment*, 11, 85–95.
- Cone, J. D. (1986). Idiographic, nomothetic and related perspectives in behavioral assessment. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 111–128). New York: Guilford Press.
- Cone, J. D. (1988). Psychometric considerations and

- the multiple models of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (pp. 42–66). New York: Pergamon.
- Cone, J. D. (1998). Psychometric considerations: Concepts, contents and methods. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Cone, J. D., & Hawkins, R. P. (Eds.) (1977). *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel.
- Craske, M. G., & Waikar, S. V. (1994). Panic disorder. In M. Hersen & R. T. Ammerman (Eds.), *Handbook of prescriptive treatments for adults*. (pp. 135–155). New York: Plenum.
- Creer, T. L., & Bender, B. G. (1993). Asthma. In R. J. Gatchel & E. B. Blanchard (Eds.), *Psychophysiological disorders, research and clinical applications* (pp. 151–204). Washington, DC: American Psychological Association.
- Davison, G. C., Navarre, S., & Vogel, R. (1995). The articulated thoughts in simulated situations paradigm: A think-aloud approach to cognitive assessment. *Current Directions in Psychological Science*, 4, 29–33.
- de Beurs, E., Van Dyck, R., van Balkom, A. J. L. M., Lange, A., & Koele, P. (1994). Assessing the clinical significance of outcome in agoraphobia research: A comparison of two approaches. *Behavior Therapy*, 25, 147–158.
- Durand, V. M., & Carr, E. G. (1991). Functional communication training to reduce challenging behavior: Maintenance and application in new settings. *Journal of Applied Behavior Analyses*, 24, 251–264.
- Durand, V. M., & Crimmins, D. M. (1988). Identifying the variables maintaining self-injurious behaviors. *Journal of Autism and Developmental Disorders*, 18, 99–117.
- Els, T. (1997). *Handbook of psychotherapy case formulation*. New York: Guilford Press.
- Eifert, G. H., & Wilson, P. H. (1991). The triple response approach to assessment: A conceptual and methodological reappraisal. *Behaviour Research and Therapy*, 29, 283–292.
- Edens, J. L., & Gil, K. M. (1995). Experimental induction of pain: Utility in the study of clinical pain. *Behavior Therapy*, 26, 197–216.
- Evans, I. (1993). Constructional perspectives in clinical assessment. *Psychological Assessment*, 5, 264–272.
- Eysenck, H. J. (1986). A critique of contemporary classification and diagnosis. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 73–98). New York: Guilford Press.
- Eysenck, H. J., & Martin, I. (1987). *Theoretical foundations of behavior therapy*. New York: Plenum.
- Figley, C. R. (Ed.) (1979), *Trauma and its wake: Volume 1: The study of post-traumatic stress disorder*. New York: Brunner/Mazel.
- Foster, S. L., Bell-Dolan, D. J. & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (pp. 119–160). New York: Pergamon.
- Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation systems. In A. R. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 253–324). New York: Wiley.
- Gannon, L. R., & Haynes, S. N. (1987). Cognitive-physiological discordance as an etiological factor in psychophysiological disorders. *Advances in Behavior Research and Therapy*, 8, 223–236.
- Gardner, W. I., & Cole, C. L. (1988). Self-monitoring procedures. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 206–246). New York: Guilford Press.
- Gatchel, R. J. (1993). Psychophysiological disorders: Past and present perspectives. In R. J. Gatchel & E. B. Blanchard (Eds.), *Psychophysiological disorders, research and clinical applications* (pp. 1–22). Washington, DC: American Psychological Association.
- Gatchel, R. J., & Blanchard, E. B. (1993). *Psychophysiological disorders, research and clinical applications*. Washington, DC: American Psychological Association.
- Glass, C. (1993). A little more about cognitive assessment. *Journal of Counseling and Development*, 71, 546–548.
- Goldfried, M. R. (1982). Behavioral Assessment: An overview. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 81–107). New York: Plenum.
- Haberman, S. J. (1978). *Analysis of qualitative data* (Vol. 1). New York: Academic Press.
- Hartmann, D. P. (Ed.) (1982). *Using observers to study behavior*. San Francisco: Jossey-Bass.
- Hatch, J. P. (1993). Headache. In: R. J. Gatchel & E. B. Blanchard (Eds.), *Psychophysiological disorders, research and clinical applications* (pp. 111–150). Washington, DC: American Psychological Association.
- Haynes, S. N. (1978). *Principles of behavioral assessment*. New York: Gardner Press.
- Haynes, S. N. (1986). The design of intervention programs. In R. O. Nelson & S. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 386–429). New York: Guilford Press.
- Haynes, S. N. (1992). *Models of causality in psychopathology: Toward synthetic, dynamic and nonlinear models of causality in psychopathology*. Des Moines, IA: Allyn & Bacon.
- Haynes, S. N. (1994). Clinical judgment and the design of behavioral intervention programs: Estimating the magnitudes of intervention effects. *Psychologia Conductual*, 2, 165–184.
- Haynes, S. N. (1996a). Behavioral assessment of adults. In M. Goldstein and M. Hersen (Eds.), *Handbook of psychological assessment*.
- Haynes, S. N. (1996b). The changing nature of behavioral assessment. In: M. Hersen & A. Bellack (Eds.), *Behavioral assessment: A practical guide* (4th ed.).
- Haynes, S. N. (1996c). The assessment–treatment relationship in behavior therapy: The role of the functional analysis. *The European Journal of Psychological Assessment*. (in press).
- Haynes, S. N., Blaine, D., & Meyer, K. (1995). Dynamical models for psychological assessment: Phase-space functions. *Psychological Assessment*, 7, 17–24.
- Haynes, S. N., Falkin, S., & Sexton-Radek, K. (1989). Psychophysiological measurement in behavior therapy. In G. Turpin (Ed.), *Handbook of clinical psychophysiology* (pp. 263–291). London: Wiley.
- Haynes, S. N., & Horn, W. F. (1982). Reactive effects of behavioral observation. *Behavioral Assessment*, 4, 369–385.
- Haynes, S. N., Leisen, M. B., & Blaine, D. D. (1997). Design of individualized behavioral treatment programs using functional analytic clinical case models. *Psychological Assessment*, 9, 334–348.
- Haynes, S. N., & O'Brien, W. O. (1990). The functional analysis in behavior therapy. *Clinical Psychology Review*, 10, 649–668.
- Haynes, S. N., & O'Brien, W. O. (1998). *Behavioral assessment. A functional approach to psychological assessment*. New York: Plenum.
- Haynes, S. N., & O'Brien, W. O. (in press). *Behavioral assessment*. New York: Plenum.
- Haynes, S. N., Spain, H., & Oliviera, J. (1993). Identifying causal relationships in clinical assessment. *Psychological Assessment*, 5, 281–291.
- Haynes, S. N., & Uchigakiuchi, P. (1993). Incorporating personality trait measures in behavioral assessment:

- Nuts in a fruitcake or raisins in a mai tai? *Behavior Modification*, 17, 72–92.
- Haynes, S. N., Uchigakiuchi, P., Meyer, K., Orimoto, Blaine, D., & O'Brien, W. O. (1993). Functional analytic causal models and the design of treatment programs: Concepts and clinical applications with childhood behavior problems. *European Journal of Psychological Assessment*, 9, 189–205.
- Haynes, S. N., & Wai'alaie, K. (1995). Psychometric foundations of behavioral assessment. In R. Fernández-Ballesteros (Ed.), *Evaluación conductual hoy: (Behavioral assessment today)* (pp. 326–356). Madrid, Spain: Ediciones Piramide.
- Haynes, S. N., & Wu-Holt, P. (1995). Methods of assessment in health psychology. In M. E. Simon (Ed.), *Handbook of health psychology* (pp. 420–444). Madrid, Spain: Sigma.
- Heatherton, T. F., & Weinberger, J. L. (Eds.) (1994). *Can personality change*. Washington, DC: American Psychological Association.
- Hersen, M., & Bellack, A. S. (Eds.) (1998). *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Hersen, M., & Turner, S. M. (Eds.) (1994). *Diagnostic interviewing* (2nd ed.). New York: Plenum.
- Hughes, H. M., & Haynes, S. N. (1978). Structured laboratory observation in the behavioral assessment of parent-child interactions: A methodological critique. *Behavior Therapy*, 9, 428–447.
- Honaker, L. M., & Fowler, R. D. (1990). Computer-assisted psychological assessment. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment*. (pp. 521–546.) New York: Pergamon.
- Iwata, B. A. (and 14 other authors). (1994). The functions of self-injurious behavior: An experimental-epidemiological analysis. *Journal of Applied Behavior Analysis*, 27, 215–240.
- Jacob, T. Tennenbaum, D., Bargiel, K., & Seilhamer, R. A. (1995). Family interaction in the home: Development of a new coding system. *Behavior Modification*, 12, 249–251.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19.
- James, L. D., Thorn, B. E., & Williams, D. A. (1993). Goal specification in cognitive-behavioral therapy for chronic headache pain. *Behavior Therapy*, 24, 305–320.
- Jensen, B. J., & Haynes, S. N. (1986). Self-report questionnaires. In A. R. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 150–175). New York: Wiley.
- Johnson, W. G., Schlundt, D. G., Barclay, D. R., Carr-Nangle, R. E., & Engler, L. B. (1995). A naturalistic functional analysis of binge eating. *Behavior Therapy*, 26, 101–118.
- Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Kanfer, F. H. (1985). Target selection for clinical change programs. *Behavioral Assessment*, 7, 7–20.
- Kazdin, A. E. (1979). Situational specificity: The two edged sword of behavioral assessment. *Behavioral Assessment*, 1, 57–75.
- Kazdin, A. (1997). *Research design in clinical psychology* (2nd ed.). New York: Allyn & Bacon.
- Kazdin, A. E., & Kagan, J. (1994). Models of dysfunction in developmental psychopathology. *Clinical Psychology: Science and Practice*, 1, 35–52.
- Kern, J. M. (1991). An evaluation of a novel role-play methodology: The standardized idiographic approach. *Behavior Therapy*, 22, 13–29.
- Kerns, R. D. (1994). Pain management. In M. Hersen & R. T. Ammerman (Eds.), *Handbook of prescriptive treatments for adults* (pp. 443–461). New York: Plenum.
- Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Erlbaum.
- Kratochwill, T. R., & Shapiro, E. S. (1988). Introduction: Conceptual foundations of behavioral assessment. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 1–13). New York: Guilford Press.
- Kubany, E. S. (1994). A cognitive model of guilt typology in combat-related PTSD. *Journal of Traumatic Stress*, 7, 3–19.
- Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50, 519–525.
- Lichstein, K. L., & Riedel, B. W. (1994). Behavioral assessment and treatment of insomnia: A review with an emphasis on clinical application. *Behavior Therapy*, 25, 659–688.
- Linscott, J., & DiGiuseppe, R. (1998). Cognitive assessment. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Malec, J. F., & Lemsky, C. (1996). Behavioral assessment in medical rehabilitation: Traditional and consensual approaches. In L. Cushman & M. Scherer (Eds.) *Psychological assessment in medical rehabilitation* (pp. 199–236). Washington, DC: American Psychological Association.
- Marsella, A. J., & Kameoka, V. (1989). Ethnocultural issues in the assessment of psychopathology. In S. Wetzler (Ed.), *Measuring mental illness: Psychometric assessment for clinicians* (pp. 157–181). Washington, DC: American Psychiatric Association.
- Mash, E. J., & Hunsley, J. (1990). Behavioral assessment: A contemporary approach. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 87–106). New York: Plenum.
- Mash, E. J., & Hunsley, J. (1993). Assessment considerations in the identification of failing psychotherapy: Bringing the negatives out of the darkroom. *Psychological Assessment*, 5, 292–301.
- Mash, E. J., & Terdal, L. G. (1988). *Behavioral assessment of childhood disorders*. New York: Guilford Press.
- McConaghy, N. (1998). Assessment of sexual dysfunction and deviation. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- McFall, R. M., & McDonel, E. (1986). The continuing search for units of analysis in psychology: Beyond persons, situations and their interactions. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 201–241). New York: Guilford Press.
- McReynolds, P. (1986). History of assessment in clinical and educational settings. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 42–80). New York: Guilford Press.
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Moran, G., Dumas, J., & Symons, D. K. (1992). Approaches to sequential analysis and the description of contingency tables in behavioral interaction. *Behavioral Assessment*, 14, 65–92.
- Munk, D. D., & Repp, A. C. (1994). Behavioral assessment of feeding problems of individuals with severe disabilities. *Journal of Applied Behavior Analysis*, 27, 241–250.
- Nelson, R. O., & Hayes, S. C. (1986). *Conceptual foundations of behavioral assessment*. New York: Guilford Press.

- Newman, M. G., Hofmann, S. G., Trabert, W., Roth, W. T., & Taylor, C. B. (1994). Does behavioral treatment of social phobia lead to cognitive changes? *Behavior Therapy*, 25, 503-517.
- Nezu, A. M., & Nezu, C. M. (1989). *Clinical decision making in behavior therapy: A problem-solving perspective*. Champaign, IL: Research Press.
- Nezu, A., Nezu, C., Friedman, & Haynes, S. N. (1996). Case formulation in behavior therapy. In T. D. Eells (Ed.), *Handbook of psychotherapy case formulation*. New York: Guilford Press.
- Nunnally, J. C., & Burnstein, I. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.
- O'Brien, W. H., & Haynes, S. N. (1995). A functional analytic approach to the conceptualization, assessment and treatment of a child with frequent migraine headaches. *In Session*, 1, 65-80.
- O'Donohue, W., & Krasner, L. (1995). *Theories of behavior therapy*. Washington, DC: American Psychological Association.
- O'Leary, D. K. (Ed.) (1987). *Assessment of marital discord: An integration for research and clinical practice*. Hillsdale, NJ: Erlbaum.
- O'Leary, K. D., Malone, J., & Tyree, A. (1994). Physical aggression in early marriage: Prerelationship and relationship effects. *Journal of Consulting and Clinical Psychology*, 62, 594-602.
- O'Leary, K. D., Vivian, D., & Malone, J. (1992). Assessment of physical aggression against women in marriage: The need for multimodal assessment. *Behavioral Assessment*, 14, 5-14.
- Ollendick, T. H., & Hersen, M. (1984). *Child behavioral assessment, principles and procedures*. New York: Pergamon.
- Ollendick, T. H., & Hersen, M. (1993). *Handbook of child and adolescent assessment*. Boston: Allyn & Bacon.
- Persons, J. B. (1989). *Cognitive therapy in practice: A case formulation approach*. New York: Norton.
- Persons, J. B., & Bertagnolli, A. (1994). Cognitive-behavioral treatment of multiple-problem patients: Application to personality disorders. *Clinical Psychology and Psychotherapy*, 1, 279-285.
- Persons, J. B., & Fresco, D. M. (1998). Assessment of depression. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD clinical psychology programs. *Journal of Personality Assessment*, 61, 394-405.
- Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. *Journal of the American Medical Association*, 264, 2511-2518.
- Rychtarik, R. G., & McGillicuddy, N. B. (1998). Assessment of appetitive disorders: Status and empirical methods in alcohol, tobacco, and other drug use. In M. Hersen & A. S. Bellack (Eds.), *Behavioral Assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Sarwer, D., & Sayers, S. L. (1998). Behavioral interviewing. In M. Hersen, & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Schmidt, N. B., & Telch, M. J. (1994). Role of fear and safety information in moderating the effects of voluntary hyperventilation. *Behavior Therapy*, 25, 197-208.
- Schlundt, D. G., Johnson, W. G., & Jarrell, M. P. (1986). A sequential analysis of environmental, behavioral, and affective variables predictive of vomiting in bulimia nervosa. *Behavioral Assessment*, 8, 253-269.
- Segal, D. L., & Fal, S. B. (1998). Structured diagnostic interviews and rating scales. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods*, 1, 47-65.
- Shapiro, E. S. (1984). Self-monitoring. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 350-373). Elmsford, NY: Pergamon.
- Shapiro, E. W., & Kratochwill, T. R. (Eds.) (1988). *Behavioral assessment in schools, Conceptual foundations and practical applications*. New York: Guilford Press.
- Shiffman, S. (1993). Assessing smoking patterns and motives. *Journal of Consulting and Clinical Psychology*, 61, 732-742.
- Silva, F. (1993). *Psychometric foundations and behavioral assessment*. Newbury Park, CA: Sage.
- Silverman, W. K., & Kurtines, W. M. (1998). Anxiety and phobic disorders: A pragmatic perspective. In M. Hersen, & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.
- Smith, G. T. (1994). Psychological expectancy as mediator of vulnerability to alcoholism. *Annals of the New York Academy of Sciences*, 708, 165-171.
- Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, 7, 300-308.
- Sobell, L. C., Toneatto, T., & Sobell, M. B. (1994). Behavioral assessment and treatment planning for alcohol, tobacco, and other drug problems: Current status with an emphasis on clinical applications. *Behavior Therapy*, 25, 523-532.
- Spector, P. E. (1992). *Summated rating scale construction: An introduction*. Beverly Hills, CA: Sage.
- Sprague, J. R., & Horner, R. H., (1992). Covariation within functional response classes: Implications for treatment of severe problem behavior. *Journal of Applied Behavior Analysis*, 25, 735-745.
- Strosahl, K. D., & Linehan, M. M. (1986). Basic issues in behavioral assessment. In A. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 12-46). New York: Wiley.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative observation data*. Hillsdale, NJ: Erlbaum.
- Sutker, P. B., & Adams, H. E. (Eds.) (1993). *Comprehensive handbook of psychopathology*. New York: Plenum.
- Taylor, C. B., & Agras, S. (1981). Assessment of phobia. In D. H. Barlow (Ed.), *Behavioral assessment of adult disorders* (pp. 280-309). New York: Guilford Press.
- Timberlake, W., & Farmer-Dougan, V. A. (1991). Reinforcement in applied settings: Figuring out ahead of time what will work. *Psychological Bulletin*, 110, 379-391.
- Torgrud, L. J., & Holborn, S. W. (1992). Developing externally valid role-play for assessment of social skills: A behavior analytic perspective. *Behavioral Assessment*, 14, 245-277.
- Turk, D. C., & Salovey, P. (Eds.) (1988). *Reasoning, inference, and judgment in clinical psychology*. New York: Free Press.
- Turkat, I. (1986). The behavioral interview. In A. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 109-149). New York: Wiley.
- Tryon, W. W. (1985). *Behavioral assessment in behavioral medicine*. New York: Springer.
- Tryon, W. W. (1991). *Activity measurement in psychology and medicine*. New York: Plenum.
- Tryon, W. (1996). Observing contingencies: Taxonomy and methods. *Clinical Psychology Review* (in press).
- Tryon, W. W. (1998). Behavioral observation. In

- M. Hersen & A. S. Bellack (Eds.). *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon. (in press)
- Waddell, G., & Turk, D. C. (1992). Clinical assessment of low back pain. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (pp. 15–36). New York: Guilford Press.
- Weiss, R. L., & Summers, K. J., (1983). Marital interaction coding system—III. In E. E. Filsinger (Ed.), *Marriage and family assessment: A sourcebook for family therapy* (pp. 85–115). Beverly Hills, CA: Sage.
- Wolpe, J., & Turkat, I. D. (1985). Behavioral formulation of clinical cases. In I. Turkat (Ed.), *Behavioral cases formulation* (pp. 213–244). New York: Plenum.

4.08

Intellectual Assessment

ALAN S. KAUFMAN

Yale University School of Medicine, New Haven, CT, USA

and

ELIZABETH O. LICHTENBERGER

The Salk Institute, La Jolla, CA, USA

4.08.1 INTRODUCTION	188
4.08.1.1 <i>Brief History of Intelligence Testing</i>	188
4.08.1.2 <i>Controversy Over Intelligence Testing</i>	190
4.08.1.3 <i>Principles of the Intelligent Testing Philosophy</i>	192
4.08.2 MEASURES OF INTELLIGENCE	193
4.08.2.1 <i>Wechsler's Scales</i>	193
4.08.2.1.1 <i>Wechsler Primary and Preschool Intelligence Scale-Revised (WPPSI-R)</i>	193
4.08.2.1.2 <i>Wechsler Intelligence Scale for Children-3rd Edition (WISC-III)</i>	195
4.08.2.1.3 <i>WISC-III Short Form</i>	198
4.08.2.1.4 <i>Wechsler Adult Intelligence Scale-Revised (WAIS-R)</i>	199
4.08.2.1.5 <i>Wechsler Adult Intelligence Scale-Third Edition (WAIS-III)</i>	203
4.08.2.1.6 <i>Kaufman Assessment Battery for Children (K-ABC)</i>	208
4.08.2.1.7 <i>Kaufman Adolescent and Adult Intelligence Test</i>	211
4.08.2.1.8 <i>Overview</i>	214
4.08.2.1.9 <i>The Stanford-Binet: Fourth edition</i>	216
4.08.2.1.10 <i>Woodcock-Johnson Psycho-Educational Battery-Revised: tests of cognitive ability (WJ-R)</i>	217
4.08.2.1.11 <i>Detroit Tests of Learning Aptitude (DTLA-3)</i>	220
4.08.2.1.12 <i>Differential Abilities Scales (DAS)</i>	221
4.08.2.1.13 <i>Cognitive Assessment System (CAS)</i>	223
4.08.3 INSTRUMENT INTEGRATION	224
4.08.3.1 <i>K-ABC Integration with Wechsler Scales</i>	224
4.08.3.2 <i>Integration of KAIT with Wechsler Scales</i>	224
4.08.3.3 <i>Integration of Binet IV with Wechsler Scales</i>	225
4.08.3.4 <i>Integration of WJ-R with Wechsler Scales</i>	225
4.08.3.5 <i>DTLA-3 Integration with Wechsler Scales</i>	226
4.08.3.6 <i>Integration of DAS with Wechsler Scales</i>	226
4.08.3.7 <i>Integration of CAS with Wechsler Scales</i>	227
4.08.4 FUTURE DIRECTIONS	227
4.08.5 SUMMARY	228
4.08.6 ILLUSTRATIVE CASE REPORT	229
4.08.6.1 <i>Referral and Background Information</i>	229
4.08.6.2 <i>Appearance and Behavioral Characteristics</i>	231
4.08.6.3 <i>Tests Administered</i>	231
4.08.6.4 <i>Test Results and Interpretation</i>	232
4.08.6.5 <i>Summary and Diagnostic Impressions</i>	233
4.08.6.6 <i>Recommendations</i>	234
4.08.7 REFERENCES	235

4.08.1 INTRODUCTION

The assessment of intellectual ability has grown and continued to develop and flourish since the nineteenth century. This chapter gives a foundation for understanding the progression of intellectual assessment through a brief historical review of IQ testing. Some of the controversy surrounding intelligence testing will be introduced, and an “intelligent” approach to testing (Kaufman, 1979, 1994b) will be discussed in response to the critics of testing. There are multiple available tests for assessing child, adolescent, and adult intelligence, and this chapter will address a select group of measures. A description and brief overview is provided on the following intelligence tests: Wechsler Primary and Preschool Intelligence Scale-Revised (WPPSI-R; Wechsler, 1989), Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1974), Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981), Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983), Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993), Stanford-Binet: Fourth Edition (Binet IV; Thorndike, Hagen, & Sattler, 1986b), Woodcock-Johnson Psycho-Educational Battery-Revised: Tests of Cognitive Ability (WJ-R; Woodcock & Johnson, 1989), Detroit Tests of Learning Aptitude (DTLA-3; Hammill, 1991), Differential Abilities Scale (DAS; Elliott, 1990), and Das-Naglieri Cognitive Assessment System (CAS; Naglieri & Das, 1996).

A thorough cognitive assessment contains supplemental measures in addition to the main instrument used to measure IQ, as will be discussed later in this chapter. Accordingly, following the description and overview of these multiple instruments is a section which integrates each of them with the Wechsler scales, focusing on how they may be used with Wechsler's tests to further define cognitive functioning. The final part of this chapter provides an illustrative case report that combines a number of different measures in the assessment of a 13-year-old female with academic difficulties.

4.08.1.1 Brief History of Intelligence Testing

Exactly when intelligence testing began is difficult to pinpoint. IQ tests as they are known in the 1990s stem from nineteenth century Europe. Study of the two extremes of intelligence, giftedness and retardation, led to breakthroughs in intellectual assessment. Tracing the roots of assessment, the early pioneers were Frenchmen, who worked with the retarded. In

the 1830s, Jean Esquirol distinguished between mental retardation and mental illness, un lumping idiocy from madness (Kaufman, 1983). He focused on language and speech patterns, and even on physical measurements such as the shape on the skull, in testing “feeble-minded” and “demented” people.

Another contribution of Esquirol was a system for labeling the retarded. He formed a hierarchy of mental retardation, with “moron” at the top. Those less mentally adept were classified as “imbecile” and those at the bottom rung of intelligence were “idiots.” The 1990s classification systems, which use terms like profoundly, severely, or moderately retarded, appear to most as less offensive than Esquirol's labels.

In the mid-1800s, another innovator joined Esquirol in testing retarded individuals. As opposed to Esquirol's use of verbal tests, Edouard Seguin tested these individuals using nonverbal methods, oriented toward sensation and motor activity (Kaufman, 1983). A link between Seguin's work and the twentieth century can be seen, as many of the procedures he developed were adopted or modified by later developers of performance and nonverbal tasks. Intelligence testing and education became intertwined during this time when Seguin convinced authorities of the desirability of educating the “idiots” and “imbeciles.” Seguin was also the inspiration for Maria Montessori. Many of his methods and materials are present in the Montessori approach to education.

In an approach similar to Seguin's, stressing discrimination and motor control, Galton studied individual differences in the ordinary man, not just the tail ends of the normal curve. He was committed to the notion that intelligence is displayed through the use of the senses (sensory discrimination and sensory motor coordination), and believed that those with the highest IQ should also have the best discriminating abilities. Therefore, he developed mental tests that were a series of objective measurements of sensory abilities like keenness of sight, color discrimination, and pitch discrimination; sensory motor-abilities like reaction time and steadiness of hand; and motor abilities like strength of squeeze and strength of pull (Cohen, Montague, Nathanson & Swerdlik, 1988).

Galton's theory of intelligence was simplistic: people take in information through their senses, so those with better developed senses ought to be more intelligent. Although his theory of intelligence was quite different than what is considered as intelligence today, he is credited with establishing the first comprehensive individual intelligence test. He also influenced two basic notions of intelligence: the idea that

intelligence is a unitary construct, and that individual differences in intelligence are largely genetically determined (possibly influenced by the theory of his cousin, Charles Darwin) (Das, Kirby, & Jarman, 1979).

Galton's concepts were brought to the USA by James McKeen Cattell, an assistant in Galton's laboratory (Roback, 1961). In 1890, Cattell established a Galton-like mental test laboratory at the University of Pennsylvania, and he moved his laboratory to Columbia University in New York the next year. He shared Galton's philosophy that intelligence is best measured by sensory tasks, but expanded his use of "mental tasks" to include standardized administration procedures. He urged for the establishment of norms, and thereby took the assessment of mental ability out of the arena of abstract philosophy and demonstrated that mental ability could be studied experimentally and practically. Studies conducted around the turn of the century at Cattell's Columbia laboratory showed that American versions of Galton's sensory-motor test correlated close to zero with meaningful criteria of intelligence, such as grade-point average in college.

Following Esquirol's lead by focusing on language abilities, Alfred Binet began to develop mental tasks with his colleagues Victor Henri and Theodore Simon (Binet & Henri, 1895; Binet & Simon, 1905, 1908). His tests were complex, measuring memory, judgment, reasoning, and social comprehension, and these tasks survive to the 1990s in most tests of intelligence for children and adults.

The Minister of Public Instruction in Paris appointed Binet to study the education of retarded children in 1904. The Minister wanted Binet to separate retarded from normal children in the public schools. Thus, with 15 years worth of task development behind him, the Binet-Simon scale quickly was assembled (Sattler, 1988).

Binet used a new approach in his tests; he ordered tasks from easy to hard within the scale. In 1908 and 1911 he revised his test to group tasks by age level, to add levels geared for adults, to introduce the concept of mental age, and to give more objective scoring rules (Sattler, 1988). If someone passed the nine-year-old level tasks, but failed the ones at the 10-year level, then that person had the intelligence of the typical nine-year old, whether the person was 6, 9, or 30. The measurement adult's intelligence, except mentally retarded individuals, was almost an after thought. Binet's untimely death in 1911 prevented him from actualizing the many applications of his tests in child development, education, medicine, and research (Kaufman, 1990).

Every IQ test in existence has been impacted greatly by Binet's work, and incorporates many of the same kinds of concepts and test questions that he developed. Lewis Terman was one of several Americans who translated the Binet-Simon for use in the USA. Terman published a "tentative" revision in 1912. Terman then took four years to carefully adapt, expand, and standardize the Binet-Simon. After much painstaking work, in 1916 the Stanford-Binet was born. This test used the concept of mental quotient and introduced the intelligence quotient. The Stanford-Binet left its competitors in the dust, and became the leading IQ test in America.

Like Binet, Terman viewed intelligence tests as useful for identifying "feeble-mindedness," or weeding out the unfit. Terman also saw the potential for using intelligence tests with adults for determining ability to perform well in certain occupations. He believed that minimum intelligence quotients were necessary for success in specific occupations.

With the onset of World War I, the field of adult assessment grew quickly due to practical recruitment issues. The USA needed a way to evaluate the mental abilities of thousands of recruits and potential officers in an expedient manner. Due to the large volume of individuals tested, a group version of Binet's test was created by Arthur Otis, a student of Terman. This group-administered Stanford-Binet was labeled the Army Alpha. The Army Beta was also created during World War I to assess anyone who could not speak English or who was suspected of malingering. This was a nonverbal problem solving test, which was a forerunner of today's nonverbal ("Performance") subtests. The Army Alpha and Army Beta tests, published by Yerkes in 1917, were validated on huge samples (nearly two million). The tests were scores "A" to "D-" with the percent scoring "A" supporting their validity: 7% of recruits, 16% of corporals, 24% of sergeants, and 64% of majors. The best evidence of validity, though, was the Peter Principle in action. Second lieutenants (59% "A") outperformed their direct superiors, first lieutenants (53%) and captains (53%), while those with ranks higher than major did not do as well as majors (Kaufman, 1990).

The subtests developed by Binet and World War I psychologists were borrowed by David Wechsler in the mid-1930s to develop the Wechsler-Bellevue Intelligence Scale. His innovation was not in the selection of tasks, but in his idea that IQ was only in part a verbal intelligence. He also assembled a Performance Scale from the nonverbal, visual-motor subtests that were developed during the war to evaluate

people who could not speak English very well or whose motivation to succeed was in doubt. Wechsler paired the verbally laden Army Alpha and the Stanford–Binet to create the verbal scale, and the Army Group Examination Beta and the Army Individual Performance scale to create the Performance scale. These two scales together were thought to contribute equally to the overall intelligence scale. The Full Scale IQ, for Wechsler, is an index of general mental ability (*g*).

To Wechsler, these tests were dynamic clinical instruments, more than tools to subdivide retarded individuals (Kaufman, 1990). However, the professional public was leery. They wondered how tests developed for the low end of the ability spectrum could be used to test normal people's intelligence. The professionals and publishers had a difficult time accepting that nonverbal tests could be used as measures for all individuals, not just foreigners. The postwar psychological community held the belief that IQ tests were primarily useful for predicting children's success in school, but were critical of Wechsler for developing a test primarily for adolescents and adults.

He persisted with his idea that people with poor verbal intelligence may be exceptional in their nonverbal ability, and vice versa. He met with resistance and frustration, and could not find a publisher willing to subsidize his new test. Thus, with a group of psychologist friends, Wechsler tested nearly 2000 children, adolescents, and adults in Brooklyn, New York. Although it was a very urban sample, he managed to obtain a well stratified sample.

Once it had been standardized, Wechsler had no problem finding a publisher in The Psychological Corporation. The original Wechsler–Bellevue (Wechsler, 1939) has grandchildren, including the Wechsler Intelligence Scale for Children–Revised (WISC-R), and the Wechsler Adult Intelligence Scale–Revised (WAIS-R); more recently in 1991 a great-grandchild was born, the WISC-III.

Loyalty to the Stanford–Binet prevented Wechsler's test from obtaining instant success. However, gradually, Wechsler overtook the Binet during the 1960s as the learning disabilities movement gained popularity. The Verbal IQ and Performance IQ provided by Wechsler's tests helped to identify bright children who had language difficulties or visual-perceptual problems. The Stanford–Binet offered just one IQ, and the test was so verbally oriented that people with exceptional nonverbal intelligence were penalized.

Terman's Stanford–Binet lost favor when revisions of the battery after his death in 1956 proved to be expedient and shortsighted. In the

1990s, Wechsler's scales have proven themselves by withstanding challenges by other test developers, including the Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983), the Kaufman Adolescent and Adult Intelligence Test (KAIT) (Kaufman & Kaufman, 1993), the Differential Abilities Scale (DAS) (Elliott, 1990), and Woodcock–Johnson (Woodcock & Johnson, 1989). These many other tests are used widely, but generally remain as alternatives or supplements to the Wechsler scales.

4.08.1.2 Controversy Over Intelligence Testing

The measurement of intelligence has long been the center of debate. In the past, critics have spoken of IQ tests as “biased,” “unfair,” and “discriminatory.” The critics' arguments in the 1990s center more around what the IQ tests truly measure, as well as how or if they should be interpreted, their relevance to intervention, and their scope. Despite the controversy, there is great interest and need for measurement of intelligence, especially in the educational context, in order to help children and adolescents. Amidst the criticisms and limitations of IQ testing, these instruments remain a most technologically advanced and sophisticated tool of the profession for providing essential and unique information to psychologists so they may best serve the needs of children and adults. When used in consideration of the American Psychological Association's Ethical Principles of Psychologists (American Psychological Association, 1990) Principle 2-Competence, which encourages clinicians to recognize differences among people (age, gender, socioeconomic, and ethnic backgrounds) and to understand test research regarding the validity and the limitations of their assessment tools, these tests can be beneficial despite the controversy.

Three controversial themes associated with IQ testing were noted by Kaufman (1994). The first involves opposition to the common practice of subtest interpretation advocated by Wechsler (1958) and Kaufman (1979, 1994b). The second includes those who would abandon the practice altogether. Finally, the third group suggests that the concept of intelligence testing is sound, but more contemporary instrumentation could improve the effectiveness of the approach.

The first group of psychologists has encouraged practitioners to “just say ‘no’ to subtest analysis” (McDermott, Fantuzzo, & Glutting, (1990) (p. 299; also see Glutting, McDermott, Prifitera, & McGrath, (1994), and Watkins & Kush, (1994)). McDermott and his colleagues argue that interpreting a subtest profile is in

violation of the principles of valid test interpretation because the ipsative method fails to improve prediction (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992) and therefore does not augment the utility of the test. It is agreed that the results of studies conducted by McDermott et al. (1992), do suggest that using the WISC-III in isolation has limitations, but using the ipsative approach in conjunction with other relevant information such as achievement test results and pertinent background information may be beneficial. Kaufman (1994) further suggests that by shifting to the child's midpoint score a more equally balanced set of hypotheses can be developed which can be integrated with other findings to either strengthen or disconfirm hypotheses. When the ipsative assessment approach is used to create a base from which to search for additional information to evaluate hypothesized strengths and weaknesses in the child's subtest profile its validity is extended beyond that which can be obtained using the Wechsler subtests alone. If support is found for the hypotheses, then such a strength or weakness can be viewed as reliable, because of its cross-validation (Kaufman, 1994). When considering this position and that represented by McDermott et al., as well as Glutting et al. (1994) and Watkins and Kush (1994), it is important to recognize that these authors are against subtest profile analysis not the use of IQ tests in general. This is in contrast to others who hold a more extreme negative view of IQ testing.

One extremist group that opposes IQ testing includes those who advocate throwing away Verbal and Performance IQs, along with the subtest profile interpretation, and finally the Full Scale IQ because they insist that all that Wechsler scales measure is *g* (MacMann & Barnett, 1994). They argue that differences between the Verbal and Performance Scales on Wechsler tests hold no meaning, that conventional intelligence tests only measure *g* (and a measure of *g* is not enough to warrant the administration of such an instrument) and that Wechsler scale data do not have instructional value. These authors fail to recognize a wealth of data that illustrates that differences between the Verbal and Performance Scales can be very important. Any clinician using intelligence tests cannot ignore the numerous studies that are available that point to significant Verbal Performance differences in patients with right-hemisphere damage (Kaufman, 1990, Chapter 9), in Hispanic and Navajo children (McShane & Cook, 1985; McShane & Plas, 1984; Naglieri, 1984), and in normal adults (Kaufman, 1990, Chapter 7). If only the Full Scale IQ is interpreted, following MacMann

and Barnett's (1994) advisement that the Verbal and Performance scales are meaningless, then it prevents the fair use of these tests with those groups who have inconsistent V-P discrepancies. Moreover, contrary to what MacMann and Barnett (1994) suggest, it is clear that when a child earns very poor Verbal and average Performance scores there are obvious implications for instruction and a high probability that such results will be reflected in poor verbal achievement (Naglieri, 1984).

Another extremist group opposed to IQ testing is Witt and Gresham (1985) who state, "The WISC-R lacks treatment validity in that its use does not enhance remedial interventions for children who show specific academic skill deficiencies" (p. 1717). It is their belief that the Wechsler test should be replaced with assessment procedures that have more treatment validity. However, as Kaufman (1994) points out, Witt and Gresham (1985) do not provide evidence for their statements. Another pair of researchers (Rechsly & Tilly, 1993) agree with the Witt and Gresham (1985) statements about the lack of treatment validity of the WISC-R, but only provide references that are not specific to the Wechsler scales. Thus, the Wechsler scales appear to have been rejected by these researchers without ample relevant data.

Witt and Gresham (1985) also complain that the WISC-R (as well as the WISC-III) only yields a score, and does not provide school psychologists with direct strategies of what to do with and for children, which are what teachers are requesting. As Kaufman (1994) points out, however, it is not the instrument's responsibility to provide direct treatment information; rather, he states, "It is the examiner's responsibility . . . to provide recommendations for intervention" (p. 35). The examiner should not be just taking the bottom-line IQ scores or standard scores, but should provide statements about a child's strengths and weaknesses that have been cross-validated through the observations of behavior, background information, and the standardized intelligence and achievement tests.

Finally, there is a group of professionals who have suggested that the Wechsler has limits that should be recognized, but these limits could be addressed by alternative methods rather than abandoning the practice of intelligence testing altogether. Some have argued for a move toward alternative conceptualizations of intelligence and methods to measure new constructs that are based on factor analytic research (e.g., Woodcock, 1990) while others have used neuropsychology and cognitive psychology as a starting point (e.g., Naglieri & Das, 1996). The results of these efforts have been tests such as the Das-Naglieri Cognitive Assessment System

(Naglieri & Das, 1996), Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983), Kaufman Adult Intelligence Test (Kaufman & Kaufman, 1993), and the Woodcock–Johnson Tests of Cognitive Ability (Woodcock & Johnson, 1989). This chapter will show how these tests, and others, can be utilized in conjunction with the Wechsler to gain a more complete view of the child.

The main goal of this chapter is to use Kaufman's (1994) philosophy of "intelligent" testing to address some of the concerns about Wechsler interpretation through a careful analysis of the results and integration with other measures. Much of this discussion is based on the principles of IQ testing as outlined by Kaufman, which focus on the view that "WISC-III assessment is of the individual, by the individual, and for the individual" (Kaufman, 1994, p. 14). Through research knowledge, theoretical sophistication, and clinical ability examiners must generate hypotheses about an individual's assets and deficits and then confirm or deny these hypotheses by exploring multiple sources of evidence. Well-validated hypotheses must then be translated into meaningful, practical recommendations. A brief description of those five principles of intelligent testing follows.

Clinician-scientists must come well equipped with state of the art instrumentation, good judgment, knowledge of psychology, and clinical training to move beyond the obtained IQs (Kaufman, 1994). Integration of information from many sources and different tests is very important if the child referred for evaluation is to remain the focus of assessment because it is impossible to describe fully a person by just presenting a few numbers from the Wechsler protocol or those obtained from a computer program. Each adult and child who comes for an assessment has unique characteristics, a particular way of approaching test items, and may be affected differently by the testing situation than the next individual. Through the use of an integrated interpretation approach the various dimensions that influence a child can become apparent.

4.08.1.3 Principles of the Intelligent Testing Philosophy

The first principle of intelligent testing is that "the WISC-III subtests measure what the individual has learned" (Kaufman, 1994, p. 6). The WISC-III is like an achievement test, in that it is a measure of past accomplishments and is predictive of success in traditional school subjects. Research indicates that intelligence tests consistently prove to be good predictors of conventional school achievement.

The WISC-III Manual (Wechsler, 1991, pp. 206–209) gives many such correlations between the WISC-III IQs or Factor Indexes and achievement measures. Although this connection between the WISC-III and achievement in school is well documented empirically, it should not be accepted ultimately as a statement of fate, that if a child scores poorly on the WISC-III that they will do poorly in school (Kaufman, 1994). Instead, constructive interpretation of a test battery can lead to recommendations which may alter helpfully a child's development.

The second principle is that WISC-III subtests are samples of behavior and are not exhaustive. Because the subtests only offer a brief glimpse into a child's overall level of functioning, examiners must be cautious in generalizing the results to performance and behaviors in other circumstances. The Full Scale "should not be interpreted as an estimate of a child's global or total intellectual functioning; and the WISC-III should be administered along with other measures, and the IQs interpreted in the context of other test scores" (Kaufman, 1994, p. 7). It is important that the actual scores are not emphasized as the bottom-line; rather, it is more beneficial to elaborate on what children can do well, relative to their own level of ability. Such information can be used to create an individualized education program which will tap a child's areas of strength and help improve areas of deficit.

Principle three states, "The WISC-III assesses mental functioning under fixed experimental conditions" (Kaufman, 1994, p. 8). Rigid adherence to the standardized procedures for administration and scoring, outlined in the WISC-III manual (Wechsler, 1991), helps to ensure that all children are measured in an objective manner. However, parts of the standardized procedure make the testing situation very different than a natural setting. For example, it is not very often in a children's every day life that someone is transcribing virtually every word they say or timing them with a stopwatch. The standardization procedures are important to follow, but must be taken into account as limitations when interpreting the scores obtained in the artificial testing situation. The value of the intelligence test is enhanced when the examiner can meaningfully relate observations of the child's behaviors in the testing situation to the profile of scores.

The fourth principle is that "The WISC-III is optimally useful when it is interpreted from an information-processing model" (Kaufman, 1994, p. 10). This is especially beneficial for helping to hypothesize functional areas of strength and dysfunction. This model suggests

examining how information enters the brain from the sense organs (input), how information is interpreted and processed (integration), how information is stored for later retrieval (storage), and how information is expressed linguistically or motorically (output). Through this model, examiners can organize the test data, including fluctuations in subtest scores, into meaningful underlying areas of asset and deficit.

The fifth and very important principle of intelligent testing is that, "Hypotheses generated from WISC-III profiles should be supported with data from multiple sources" (Kaufman, 1994, p. 13). Although good hypotheses can be raised from the initial WISC-III test scores, such hypotheses do not hold water unless verified by diverse pieces of data. Such supporting evidence may come from careful observation of a child's behavior during test administration; from the pattern of responses across various subtests; from background information obtained from parents, teachers, or other referral sources; from previous test data; and from the administration of supplemental subtests. The integration of data from all these different sources is critical in obtaining the best and most meaningful clinical interpretation of a test battery.

4.08.2 MEASURES OF INTELLIGENCE

Intelligence tests are administered for a variety of reasons including identification (of mental retardation, learning disabilities, other cognitive disorders, giftedness), placement (gifted and other specialized programs), and as a cognitive adjunct to a clinical evaluation. The following comprehensive intelligence tests are discussed in the next sections: WPPSI-R, WISC-III, WAIS-R, K-ABC, KAIT, Binet-IV, WJ-R Tests of Cognitive Ability, DTLA-3, DAS, and CAS.

4.08.2.1 Wechsler's Scales

As discussed in the brief history of IQ tests, Wechsler's scales reign as leaders of measures of child, adolescent, and adult intelligence. The WISC-III is a standard part of a battery administered to children by school psychologists and private psychologists to assess level cognitive functioning, learning styles, learning disabilities, or giftedness. The WAIS-R is administered invariably as a part of a battery to assess intellectual ability for a clinical, neurological, or vocational evaluation of adolescents and adults. The WPPSI-R may be used to measure intellectual ability from ages three to seven years, three months; intellectual assessment may be done from age six up to age 16 with

the WISC-III, while the WAIS-R may be used from ages 16–74.

The different Wechsler scales overlap at ages 6–7 and 16. Kaufman (1994) recommends that the WISC-III be used at both these overlapping age periods rather than the WPPSI-R or the WAIS-R. One of the reasons cited for these recommendations is that the WISC-III has a much better "top" than the WPPSI-R for children who are ages six or seven. On the WPPSI-R a child can earn a maximum score of 16 or 17 (rather than 19) on six of the 10 subtests when age seven. The recommendation to use the WISC-III rather than the WAIS-R at age 16 is made because the WAIS-R norms are outdated relative to the WISC-III norms. Kaufman (1990) recommends that the WAIS-R norms for ages 16–19 be used cautiously, and states that the "even apart WISC-III norms, the Performance scale is more reliable for the WISC-III (0.92) than the WAIS-R (0.88) at age 16" (p. 40).

Wechsler (1974) puts forth the definition that "intelligence is the overall capacity of an individual to understand and cope with the world around him [or her]" (p. 5). His tests, however, were not predicated on this definition. Tasks developed were not designed from well-researched concepts exemplifying his definition. In fact, as previously noted, virtually all of his tasks were adapted from other existing tests.

Like the Binet, Wechsler's definition of intelligence also ascribes to the conception of intelligence as an overall global entity. He believed that intelligence cannot be tested directly, but can only be inferred from how an individual thinks, talks, moves, and reacts to different stimuli. Therefore, Wechsler did not give credence to one task above another, but believed that this global entity called intelligence could be ferreted out by probing a person with as many different kinds of mental tasks as one can conjure up. Wechsler did not believe in a cognitive hierarchy for his tasks, and he did not believe that each task was equally effective. He felt that each task was necessary for the fuller appraisal of intelligence.

4.08.2.1.1 Wechsler Primary and Preschool Intelligence Scale-Revised (WPPSI-R)

(i) Standardization and properties of the scale

The WPPSI-R is an intelligence test for children aged three years, through seven years, three months. The original version of the WPPSI was developed in 1967 for ages four to six and a half years, and the WPPSI-R was revised in 1989. Several changes were made to the revised version of the WPPSI-R. The norms were updated, the appeal of the content to

young children was improved, and the age range was expanded.

The WPPSI-R is based on the same Wechsler–Bellevue theory of intelligence, emphasizing intelligence as a global capacity but having Verbal and Performance scales as two methods of assessing this global capacity (Kamphaus, 1993). The Verbal scale subtests include: Information Comprehension, Arithmetic, Vocabulary, Similarities, and Sentences (optional subtest). The Performance scale subtests include: Object Assembly, Block Design, Mazes, Picture Completion, and Animal Pegs (optional subtest).

Like the K-ABC and the Differential Abilities Scales (DAS), the WPPSI-R allows the examiner to “help” or “teach” the client on early items on the subtests to assure that the child understands what is expected. Providing this extra help is essential when working with reticent preschoolers (Kamphaus, 1993).

Subtest scores have a mean of 10 and a standard deviation of three. The overall Verbal, Performance, and Full Scale IQs have a mean of 100 and a standard deviation of 15. The examiner manual provides interpretive tables that allow the examiner to determine individual strengths and weaknesses as well as the statistical significance and clinical rarity of Verbal and Performance score differences.

The WPPSI-R was standardized on 1700 children from age three through seven years, 3 months. The standardization procedures followed the 1986 US Census Bureau estimates. Stratification variables included gender, race, geographic region, parental occupation, and parental education.

The WPPSI-R appears to be a highly reliable measure. The internal consistency coefficients across age groups, for the Verbal, Performance, and Full Scale IQs are 0.95, 0.92, and 0.96, respectively. For the seven-year-old age group, the reliability coefficients are somewhat lower. The internal consistency coefficients for the individual Performance subtests vary from 0.63 for Object Assembly to 0.85 for Block Design, with a median coefficient of 0.79. The internal consistency coefficients for the individual Verbal subtests vary from 0.80 for Arithmetic to 0.86 for Similarities, with a median coefficient of 0.84. The test–retest coefficient for the Full Scale IQ is 0.91.

The WPPSI-R manual provides some information on validity; however, it provides no information on the predictive validity of the test. Various studies have shown that concurrent validity between the WPPSI-R and other tests is adequate. The correlation between the WPPSI and the WPPSI-R Full Scale IQs was reported at 0.87, and the correlations between WPPSI-R

and WISC-III Performance, Verbal, and Full Scale IQs for a sample of 188 children were 0.73, 0.85, and 0.85, respectively. The correlations between the WPPSI-R and other well known cognitive measures are, on average, much lower. The WPPSI-R Full Scale IQ correlated 0.55 with the K-ABC Mental Processing Composite (Kamphaus, 1993) and 0.77 with the Binet IV Test Composite (McCrowell & Nagle, 1994). In general, the validity coefficients provide strong evidence for the construct validity of the WPPSI-R (Kamphaus, 1993).

(ii) Overview

The WPPSI-R is a thorough revision of the 1967 WPPSI, with an expanded age range, new colorful materials, new item types for very young children, a new icebreaker subtests (Object Assembly), and a comprehensive manual (Kaufman, 1990). The revision of the test has resulted in an instrument that is more attractive, more engaging, and materials that are easier to use (Buckhalt, 1991; Delugach, 1991).

The normative sample is large, provides recent norms and is representative of the 1986 US Census data (Delugach, 1991; Kaufman, 1990). The split-half reliability of the IQs and most subtests are exceptional, the factor analytic results for all age groups are excellent, and the concurrent validity of the battery is well supported by several excellent correlational studies (Delugach, 1991; Kaufman, 1990). The manual provides a number of validity studies, factor analytic results, research overviews, and state-of-the-art interpretive tables, which provide the examiner with a wealth of information. Kaufman (1990) noted that the WPPSI-R has a solid psychometric foundation.

In spite of its reported strengths, the WPPSI-R has flaws. In publishing the WPPSI-R, great effort was made to ensure that all subtests had an adequate “top” and “bottom” (Kaufman, 1992). However, the WPPSI-R has an insufficient floor at the lowest age levels, which limits the test’s ability to diagnose intellectual deficiency in young preschoolers (Delugach, 1991). For example, a child at the lowest age level (2–11–16 to 3–2–15) who earns only one point of credit on all subtests will obtain a Verbal IQ of 75, a Performance IQ of 68, and a Full Scale IQ of 68, making it impossible to adequately assess the child’s degree of intellectual deficiency. The WPPSI-R subtests are able to distinguish between gifted and nongifted children at all ages, but the top of some subtests is not adequate to discriminate among gifted children. Kaufman (1992) indicates that at the youngest ages (3–4.5 years), all subtests are excellent. However, at age five, Geometric Design begins

to falter, and at ages 6.5 and above, it only allows a maximum scaled score of 16. Other subtests, such as Object Assembly and Arithmetic also have problems with the ceiling. Although the ceilings on the subtests described are not ideal, the IQ scales do allow maximum IQs of 150 for all ages and IQs of 160 for children ages 3–6.25.

Another major problem with the WPPSI-R is the role played by speed of responding. From both early developmental perspectives and common-sense perspectives, giving bonus points for speed is silly (Kaufman, 1992). Young children may respond slowly for a variety of reasons that have little to do with intellect. A three- or four-year-old child might respond slowly or deliberately because of lack of immaturity, lack of experience in test taking, underdeveloped motor coordination, or a reflective cognitive style. The WPPSI-R Object Assembly and Block Design place an over-emphasis on speed. For example, if a six and a half or seven year old child solves every Object Assembly item perfectly, but does not work quickly enough to earn bonus points, they would only receive a scaled score of 6 (ninth percentile). Because of the age-inappropriate stress on solving problems in with great speed, a child's IQ may suffer on two of the 10 subtests (Kaufman, 1992). In addition, the directions on some of the Performance subtests are not suitable for young children because they are not developmentally appropriate (Kaufman, 1990). However, Delugach (1991) reports that if the directions are too difficult, the test provides procedures to ensure that the child understands the demands of the task.

The WPPSI-R is a useful assessment tool, but, like all others, it possesses certain weaknesses that limit its usefulness (Delugach, 1991). Examiners should be aware of the WPPSI-R's inherent strengths and weaknesses and keep them in mind during administration, scoring, and interpretation. The WPPSI-R may provide the examiner with useful information; however, "it does little to advance our basic understanding of the development and differentiation of intelligence or our understanding of the nature of individual differences in intelligence" (Buckhalt, 1991).

4.08.2.1.2 Wechsler Intelligence Scale for Children-3rd Edition (WISC-III)

(i) Standardization and properties of the scale

The WISC-III was standardized on 2200 children ranging in age from six through 16 years. The children were divided into 11 age groups, one group for each year from six

through 16 years of age. The median age for the each age group was the sixth month (e.g., 7 years, 6 months). The standardization procedures followed the 1980 US Census data and the manual provides information by age, gender, race/ethnicity, geographic region, and parent education. "Overall, the standardization of the WISC-III is immaculate ... a better-standardized intelligence test does not exist" (Kaufman, 1993, p. 351).

The WISC-III yields three IQs, a Verbal Scale IQ, a Performance Scale IQ, and a Full Scale IQ. All three are standard scores (mean of 100 and standard deviation of 15) obtained by comparing an individual's score with those earned by the representative sample of age peers. Within the WISC-III, there are 10 mandatory and three supplementary subtests all of which span the age range of six through 16 years. The Verbal scale's five mandatory subtests include: Information, Similarities, Arithmetic, Vocabulary, and Comprehension. The supplementary subtest on the Verbal Scale is Digit Span. Digit Span is not calculated into the Verbal IQ unless it has been substituted for another Verbal subtest because one of those subtests has been spoiled (Kamphaus, 1993; Wechsler, 1991).

The five mandatory Performance scale's subtests include Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Coding. The two supplementary subtest on the Performance scale are Mazes and Symbol Search. The Mazes subtest may be substituted for any Performance scale subtest; however, Symbol Search may only be substituted for the Coding subtest (Kamphaus, 1993; Wechsler, 1991).

"Symbol Search is an excellent task that should have been included among the five regular Performance subtests instead of Coding. Mazes is an awful task that should have been dropped completely from the WISC-III" (Kaufman, 1994, p. 58). He goes further to say that "there's no rational reason for the publisher to have rigidly clung to Coding as a regular part of the WISC-III when the new Symbol Search task is clearly a better choice for psychometric reasons" (p. 59). Therefore, for all general purposes, Kaufman (1994) strongly recommends that Symbol Search be substituted routinely for coding as part of the regular battery, and to use Symbol Search to compute the Performance IQ and Full Scale IQ. The manual does not say to do this, but neither does it prohibit it.

Reliability of each subtest except Coding and Symbol Search was estimated by the split-half method. Stability coefficients were used as reliability estimates for the Coding and Symbol Search subtests because of their speeded nature.

Across the age groups, the reliability coefficients range from 0.69 to 0.87 for the individual subtests. The average reliability, across the age groups, for the IQs and Indexes are: 0.95 for the Verbal IQ, 0.91 for the Performance IQ, 0.96 for the Full Scale IQ, 0.94 for the Verbal Comprehension Index, 0.90 for the Perceptual Organization Index, 0.87 for the Freedom from Distractibility Index and 0.85 for the Processing Speed Index (Wechsler, 1991).

Factor analytic studies of the WISC-III standardization data were performed for four age group subsamples: ages 6–7 ($n = 400$), ages 8–10 ($n = 600$), ages 11–13 ($n = 600$), and ages 14–16 ($n = 600$) (Wechsler, 1991). Compiling the results of the analysis, a four-factor solution was found for the WISC-III. Like the WISC-R, Verbal Comprehension and Perceptual Organization remain the first two factors. Verbal Comprehension involves verbal knowledge and the expression of this knowledge. Perceptual Organization, a nonverbal dimension, involves the ability to interpret and organize visually presented material. The third factor consists of the Arithmetic and Digit Span subtests. Factor III has been described as Freedom from Distractibility since common among tasks is the ability to focus, to concentrate, and to remain attentive. Other interpretations of this factor have included facility with numbers, short-term memory, and sequencing because the three tasks which comprise the factor all involve a linear process whereby numbers are manipulated. Success is either facilitated by or wholly dependent on memory (Kaufman, 1979). The fourth factor consists of Coding and Symbol Search, and is referred to as the Processing Speed factor. Taken together, the Verbal Comprehension and Perceptual Organization factors offer strong support for the construct validity of the Verbal and Performance IQs; substantial loadings on the large, unrotated first factor (g) supports the construct underlying Wechsler's Full Scale IQ.

(ii) Analyzing the WISC-III data

To obtain the most information from the WISC-III, the psychologist should be more than familiar with each of the subtests individually as well as with the potential information that those subtests can provide when integrated or combined. The WISC-III is maximally useful when tasks are grouped and regrouped to uncover a child's strong and weak areas of functioning, so long as these hypothesized assets and deficits are verified by multiple sources of information.

As indicated previously, the WISC-III provides examiners with a set of four Factor Indexes in addition to the set of three IQs. The

front page of the WISC-III record form lists the seven standard scores in a box on the top right. The record form is quite uniform and laid out nicely; however, it is difficult to know just what to do with all of those scores. Kaufman (1994) has developed seven steps to interpretation which offer a systematic method of WISC-III interpretation that allows the clinician to organize and integrate the test results in a step-wise fashion. The seven steps (see Table 1) provide an empirical framework for profile attack while organizing the profile information into hierarchies.

(iii) Overview

Professionals in the field of intelligence testing have described the third edition of the Wechsler Intelligence Scale for Children in a number of different ways. Some critics feel that the WISC-III reports continuity, the status quo, but makes little progress in the evolution of the assessment of intelligence. Such critics note that despite more than 50 years of advancement in theories of intelligence, the Wechsler philosophy of intelligence (not actually a formal theory), written in 1939, remains the guiding principle of the WISC-III (Schaw, Swerdilik, & Laurent, 1993). One of the principal goals for developing the WISC-III stated in the manual was merely to update the norms, which is "hardly a revision at all" (Sternberg, 1993). Sternberg (1993) suggests that the WISC-III is being used to look for a test of new constructs in intelligence, or merely a new test, the examiner should look elsewhere.

In contrast to these fairly negative evaluations, Kaufman (1993) reports that the WISC-III is a substantial revision of the WISC-R and that the changes that have been made are considerable and well done. "The normative sample is exemplary, and the entire psychometric approach to test development, validation, and interpretation reflects sophisticated, state-of-the-art knowledge and competence" (Kaufman, 1993). For Kaufman, the WISC-III is not without its flaws but his overall review of the test is quite positive. One of Kaufman's (1993) main criticisms is that the Verbal tasks are highly culturally-saturated and school-related, which tend to penalize bilingual, minority, and learning-disabled children. He suggests that perhaps a special scale could have been developed to provide a fairer evaluation of the intelligence of children who are from the non-dominant culture or who have academic difficulties. Another criticism raised by Kaufman is that too much emphasis is placed on a child's speed of responding on the WISC-III. It is difficult to do well on the WISC-III if you do not solve problems very quickly. This need for

Table 1 Summary of seven steps for interpreting WISC-III profiles.

Step	
1	<i>Interpret the full scale IQ</i> Convert it to an ability level and percentile rank and band it with error, preferable a 90% confidence interval (about ± 5 points)
2	<i>Determine if the verbal-performance IQ discrepancy is statistically significant</i> Overall values for V-P discrepancies are <i>11 points</i> at the 0.05 level and <i>15 points</i> at the 0.01 level. For most testing purposes, the 0.05 level is adequate
3	Determine if the V-P IQ discrepancy is interpretable, or if the VC and PO factor indexes should be interpreted instead Ask four questions about the Verbal and Performance Scales <i>Verbal Scale</i> (i) Is there a significant difference ($p < 0.05$) between the child's standard scores in VC vs. FD? size needed for significant (VC-FD) = 13+ points (ii) Is there abnormal scatter (highest minus lowest scaled score) among the five Verbal subtests used to compute V-IQ? Size needed for abnormal verbal scatter = 7+ points <i>Performance Scale</i> (iii) Is there a significant difference ($p < 0.05$) between the child's standard scores on PO vs. PS? Size needed for significant (PO-PS) = 15+ points (iv) Is there abnormal scatter (highest minus lowest scaled score) among the five Performance subtests used to compute P-IQ? Size needed for abnormal performance scatter = 9+ points If all answers are no, the V-P IQ discrepancy is interpretable. If the answer to one or more questions is yes, the V-P IQ discrepancy may not be interpretable. Examine the VC-PO discrepancy. Overall values for VC-PO discrepancies are <i>12 points</i> at the 0.05 level and <i>16 points</i> at the 0.01 level Determine if the VC and PO indexes are unitary dimensions: 1. Is there abnormal scatter among the four VC subtests? Size needed for abnormal VC scatter = 7+ points 2. Is there abnormal scatter among the four PO subtests? Size needed for abnormal PO scatter = 8+ points If the answer to either question is yes, then you probably shouldn't interpret the VC-PO Index discrepancy – unless the discrepancy is to big to ignore (see Step 4). If both answers are no, interpret the VC-PO differences as meaningful
4	<i>Determine if the V-P IQ discrepancy (or VC-PO discrepancy) is abnormally large</i> Differences of at least <i>19 points</i> are unusually large for both the V-P and VC-PO discrepancies. Enter the table with the IQs or Indexes, whichever was identified by the questions and answers in Step 3 If neither set of scores was found to be interpretable in Step 3, they may be interpreted anyway if the magnitude of the discrepancy is unusually large (19+ points)
5	<i>Interpret the meaning of the global verbal and nonverbal dimensions and the meaning of the small factors</i> Study the information and procedures presented in Chapter 4 (verbal/nonverbal) and Chapter 5 (FD and PS factors). Chapter 5 provides the following rules regarding when the FD and PS factors have too much scatter to permit meaningful interpretation of their respective Indexes (both Chapters 4 and 5 are on <i>Intelligent Testing with the WISC-III</i>): (i) Do not interpret the FD Index if the Arithmetic and Digit Span scaled scores differ by <i>four or more points</i> (ii) Do not interpret the PO Index if the Symbol Search and Coding scaled scores differ by <i>four or more points</i>
6	<i>Interpret significant strengths and weaknesses in the WISC-III subtest profile</i> If the V-P IQ discrepancy is less than 19 points, use the child's mean of all WISC-III subtests administered as the child's midpoint If the V-P IQ discrepancy is 19 or more points, use the child's mean of all Verbal subtests as the midpoint for determining strengths and weaknesses on Verbal subtests, and use the Performance mean for determining significant deviations on Performance subtests Use either the specific values in Table 3.3 of <i>Intelligent Testing with the WISC-III</i>, rounded to the nearest whole number, or the following summary information for determining significant deviations: ± 3 points: Information, similarities, arithmetic, vocabulary ± 4 points: Comprehension, digit span, picture completion, picture arrangement, block design, object assembly, symbol search ± 5 points: Coding
7	<i>Generate hypotheses about the fluctuations in the WISC-III subtest profile</i> Consult Chapter 6 in <i>Intelligent Testing with the WISC-III</i>, as it deals with the systematic reorganization of subtest profiles to generate hypotheses about strengths and weaknesses

speed penalizes children who are more reflective in their cognitive style or who have coordination difficulties. The speed factor may prevent a gifted child from earning a high enough score to enter into an enrichment class or may lower a learning disabled child's overall IQ score to a below average level, just because they do not work quickly enough. Although the WISC-III clearly has had mixed reviews, it is one of the most frequently used tests in the field of children's intelligence testing.

4.08.2.1.3 WISC-III Short Form

Short forms of the Wechsler scales were developed shortly after the original tests were developed (Kaufman, Kaufman, Balgopal, & McLean, 1996). Short forms are designed to have sound psychometric qualities and clinical justification, but should also be practical to use. Clinicians and researchers utilize short form when wanting to perform a screen of intellectual ability or when doing research which does not permit the needed time to complete an entire Wechsler scale. In a study using three different WISC-III short forms, the clinical, psychometric, and practical qualities for each form were examined (Kaufman, Kaufman, Balgopal et al. 1996). A psychometric and clinically strong short form was examined, and included the following subtests: Similarities, Vocabulary, Picture Arrangement, and Block Design. A practical short form, based on its brevity and ease of scoring, included the following subtests: Information, Arithmetic, Picture Completion, and Symbol Search. A short form which combines psychometric, clinical, and practical qualities was also examined: Similarities, Arithmetic, Picture Completion, and Block Design.

The results of this study using the WISC-III standardization sample of 2200 children, 6–16 years old, revealed important information about the utility of these three different short forms (Kaufman, Kaufman, Balgopal et al., 1996). The split-half reliability coefficients, standard error of measurement (SEM), validity coefficients, and standard errors of estimate for the three selected tetrads are presented in Table 2. The form which had both psychometric and clinical properties (S-V-PA-BD) was compared to the form which had the quality of practicality in addition to psychometric and clinical properties (S-A-PC-BD). The results indicated that they were equally valid and about equally reliable (see Table 2). Each of the three short form tetrads had reliability coefficients of above 0.90 for the total sample. The brief tetrad (I-A-PC-SS) had a lower correlation with the Full Scale of 0.89, compared to the other two forms which each correlated 0.93 with the Full Scale.

Although they were each about equally reliable, the S-A-PC-BD subtests are quicker to administer and only Similarities requires some subjectivity to score. It is quicker to score than the S-V-PA-BD form because it uses Arithmetic instead of Vocabulary.

The authors recommend that the extra 25–30% savings in time in using the S-A-PC-BD form, in addition to the added validity in comparison to the practical tetrad (I-A-PC-SS), makes the S-A-PC-BD short form an excellent choice. The very brief practical form was not recommended for clinical use or screening purposes because of its lower validity. Kaufman, Kaufman, Balgopal, et al. (1996) present an equation for converting a person's sum of scaled scores of the four subtests to estimated FSIQs. The magnitude of the intercorrelations among the component subtests provide that data from which the exact equation is derived. The intercorrelations vary to some extent as a function of age, which leads to slightly different equations at different ages. However, the authors state that the equations for the total sample for each tetrad represent a good overview for all children ages 6–16. The following conversion equation is for the total sample for the recommended S-A-PC-BD short form:

Estimated FSIQ = $1.6c + 36$, (for other specific equations for varying ages, see Kaufman, Kaufman, Balgopal et al., 1996, p. 103).

To use this conversion equation, the child's scaled scores on the four subtests (S-A-PC-BD) must first be summed. The sum (X_c) must then be entered into the equation. For example, if examiners give the recommended psychometric/clinical/practical form to an eight-year-old, the child's scores on the four subtests would need to be summed. Suppose that the child's sum is 50. The above equation would show:

$$\begin{aligned}\text{Estimated FSIQ} &= 1.6(50) + 36 \\ &= 80 + 36 = 116\end{aligned}$$

It is important to note that examiners should not take the good psychometric qualities of the short form to mean that the short form can be regularly substituted for the complete battery. There is a wealth of information, both clinical and psychometric, that the examiner benefits from when administering the complete battery. It is important not to favor the short forms just because they are shorter, on account of all the important information that is derived from a complete administration. Kaufman (1990) suggests that the following are a few instances in which the savings in administration time may justify use of the short form: (i) when only a

Table 2 Reliability, standard error of measurement, and validity of the three selected short forms by age.

Age (Years)	Split-half reliability ^a			Standard error of measurement			Validity: Correlation with full scale ^a			Standard error of estimate		
	SF1	SF2	SF3	SF1	SF2	SF3	SF1	SF2	SF3	SF1	SF2	SF3
6	92	92	90	4.2	4.2	4.7	92	94	89	6.0	5.2	7.0
7	91	90	89	4.5	4.7	5.0	91	91	89	6.4	6.4	7.0
8	93	93	92	4.0	4.0	5.5	92	94	90	6.0	5.2	6.7
9	91	91	89	4.5	4.5	5.0	93	93	89	5.6	5.6	7.0
10	93	92	90	4.0	4.2	4.7	91	89	87	6.4	4.2	7.6
11	91	92	91	4.5	4.2	4.5	94	91	89	5.2	6.4	7.0
12	94	92	90	3.7	4.2	4.7	94	93	89	5.2	5.6	7.0
13	93	92	90	4.0	4.2	4.7	94	92	86	5.2	6.0	7.9
14	94	92	90	3.7	4.2	4.7	93	93	90	5.6	5.6	6.7
15	94	94	93	3.7	3.7	4.0	96	94	91	4.2	5.2	6.4
16	94	93	92	3.7	4.0	4.2	94	95	90	5.2	4.7	6.7
Total	93	92	91	4.0	4.2	4.5	93	93	89	5.6	5.6	7.0

Source: Kaufman et al. (1996).

Notes: SF1 = Short Form 1 (Psychometric/Clinical; Similarities-Vocabulary-Picture Arrangement-Block Design), SF2 = Short Form 2 (Psychometric/Clinical/Practical; Similarities-Arithmetic-Picture Completion-Block Design), SF3 = Short Form 3 (Practical; Information-Arithmetic-Picture Completion-Symbol Search).

^a Decimal points have been omitted.

global assessment of IQ is needed in the context of a complete personality evaluation; (ii) when a thorough evaluation has been completed recently and just a brief check of present intellectual functioning is needed; and (iii) when an individual does not require categorization of their intellectual ability for placement or for diagnosis of a cognitive disorder.

4.08.2.1.4 Wechsler Adult Intelligence Scale-Revised (WAIS-R)

The Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; The Psychological Corporation, 1997) came out in August 1997 and soon will be replacing the WAIS-R. Based on our experience with previous versions of the WAIS, it is likely that there will be a transition period of 3–4 years during which time clinicians will be gradually moving to use primarily the newer instrument. Because of this predicted transition time, we are including information about both the WAIS-R and the WAIS-III. Additionally, much of the research on the WAIS-R will be directly relevant and applicable to the WAIS-III and is therefore included here.

(i) Standardization and properties of the scale

Similar to the other Wechsler scales discussed, three IQ scores are derived from the WAIS-R subtests. Each of these scores are standard scores with a mean of 100 and a standard deviation of 15, which are created by comparing an individual's score to scores earned by the normative

group of the same age. The Verbal IQ is comprised of six verbal subtest scores (Information, Digit Span, Vocabulary, Arithmetic, Comprehension, and Similarities). The Performance IQ is comprised of five nonverbal subtests (Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Digit Symbol). The Full Scale represents the average of the Verbal and Performance IQs.

The WAIS-R was standardized by administering the full scale to 1880 adult subjects, selected according to current US Census data tested between 1976 and 1980. Subjects were stratified according to age, sex, race (white–nonwhite), geographic region, occupation, education, and urban–rural residence. Subjects were divided into nine age groups, corresponding to categories often used by the US Census Bureau. The number in each age group ranged from 160–300, and the age groups spanned from ages 16–74. A problem with the standardization sample noted by Kaufman (1985) was that there was apparent systematic bias in the selection of 16- to 19-year-olds, leading to very questionable teenage norms. Also Hispanics were not included systematically in the total sample. There was no mention of how Hispanics were categorized, if any were tested (Kaufman, 1985).

Reliability coefficients for the 11 tests and the Verbal, Performance, and Full Scale IQs were computed using the split-half method (except for Digit Span and Digit Symbol). Average reliability, across the nine age groups, are as follows: 0.97 for the Verbal IQ; 0.93 for the Performance IQ; and 0.97 for the Full Scale IQ.

Stability coefficients for the three IQs are: 0.97, 0.90, and 0.96 for Verbal, Performance, and Full Scale, respectively.

Many factor analytic studies are available which examine the underlying structure of the WAIS (i.e., *WAIS-R manual*, 1981). Three basic factors have been reported: a “verbal comprehension” factor, a “perceptual organization” factor, and a “memory” factor, which has also been assigned labels like Freedom from Distractibility, Sequential Ability, and Number Ability. These findings are noted to confirm the appropriateness of separating the tests of the WAIS into the Verbal and Performance Scales. Researchers have disagreed about how many factors do underlie the WAIS-R. Some researchers regard the WAIS-R as a one-factor test, stating that the common ability factors account for only a small measure of intellectual ability (O’Grady, 1983). Some have interpreted as many as four or five meaningful WAIS factors for various normal and clinical samples (Cohen, 1957). However, Kaufman (1990) states that there is not any justification for interpreting five WAIS-R factors.

When only two factors are rotated for the WAIS-R, the results do not quite correspond to Wechsler’s division of subtests into the Verbal and Performance Scales, although the fit is adequate. In a comparison of six cross-validation samples and the total normative sample using two-factor solutions, all of the Verbal subtests loaded more highly on the Verbal Comprehension than the Perceptual Organization factor (Kaufman, 1990). The loadings from the standardization sample ranged from 0.47 to 0.84 for the Verbal Conceptualization factor, and ranged from 0.45 to 0.72 for the Perceptual Organization factor. Two Verbal tests (Digit Span and Arithmetic) did, however, show strong secondary loadings on the Performance factor. Digit Span and Arithmetic’s loadings on the Verbal dimension are also not as consistently strong as the other four Verbal Subtests. Each of the five Performance subtests also consistently loaded more highly on the Perceptual Organization than the Verbal Comprehension factor for the total standardization sample and for the various supplementary groups. Picture Arrangement was the exception, with equal loadings on both factors of the normative group (Kaufman, 1990).

The three-factor solutions for the normal WAIS-R standardization sample demonstrated factors that were fairly well anticipated. Kaufman (1990) discusses the three factor solutions, and presents a table summarizing the data from six samples plus the normative sample (p. 244). The Verbal Comprehension factor was defined by loadings ranging from 0.67 to 0.81, includ-

ing: Information, Vocabulary, Comprehension, and Similarities. The triad of Picture Completion, Block Design, and Object Assembly comprised the Perceptual Organization factor with loadings in the 0.56–0.73 range. The third factor comprises Digit Span and Arithmetic, with factors of 0.64 and 0.55, respectively. Picture Arrangement and Digit Symbol are more or less unaccounted for in the three-factor solution. Picture Arrangement loads equally on the verbal and nonverbal dimensions. Digit Symbol achieves loadings of only 0.32, 0.38, and 0.36 for each of the factors, loading only marginally on each, but not definitively on any. Depending on the profile obtained by any given individual, examiners may choose to interpret either two or three factors (Kaufman, 1990). The decision to interpret two or three factors should be based on whether the small third dimension is interpretable for a given person.

Studies on gender differences on the WAIS-R have shown that males’ earned IQs were higher (although not significantly so) than females’ earned IQs (Kaufman, 1990). In a sample of 940 males and 940 females, males scored about two points higher on the VIQ, 1.5 points higher on the PIQ, and two points higher on the FSIQ (Reynolds, Chastain, Kaufman, & McLean, 1987). When the gender differences are examined within different age groups, there are larger differences for ages 20–54 than at the extreme age groups of 16–19 and 55–74. For the 20–54 year age range, males scored higher by about 2.5 points on VIQ and PIQ, and by about three points on the FSIQ (Kaufman, 1990).

In examining gender differences on the individual subtests (Kaufman, McLean, & Reynolds, 1988), males and females were found to perform differently on some of the 11 subtests. On Information, Arithmetic, and Block Design males significantly and consistently outperformed females. However, females were far superior on Digit Symbol. On a less consistent basis, males showed superiority on Comprehension, Picture Completion, Picture Arrangement, and Object Assembly. No gender differences for any age group were found for Digit Span, Vocabulary, and Similarities.

Research on WAIS-R profiles has also focused on the area of neuropsychology. In this area it has been hypothesized that lesions in the left cerebral hemisphere are associated with diminished language and verbal abilities, whereas lesions in the right cerebral hemisphere are accompanied by visual-spatial deficits (Reitan, 1955). The hypothesis that has grown from these expected differences is that individuals with left brain lesions will have WAIS-R profiles with $P > V$, and individuals with right

hemisphere lesions will have a profile with $V > P$ (Kaufman, 1990). On the basis of numerous WAIS and WAIS-R studies of patients with brain lesions, two general conclusions may be drawn. A summary of empirical data (Kaufman, 1990) leads to a few main conclusions, as follows (see Table 3). First, patients with right hemisphere damage (unilateral or some bilateral damage as well) will most likely demonstrate a $V > P$ profile. Second, patients with left hemisphere, unilateral damage may show a slight $P > V$ profile, but not large enough in size or consistently enough that it is beneficial diagnostically.

A further area of study in subjects with unilateral brain lesions and cognitive ability is with gender differences. Males and females are believed to differ in various aspects of brain functioning. Kaufman (1990) presents data from eight studies that included males and females with brain lesions. The accumulated data are reported to support the alleged gender-related interaction between side of brain lesion and direction of Verbal IQ–Performance IQ difference. Damage to the right hemisphere for both males and females lead to more striking V – P differences than damage to the left hemisphere. However, the $V > P$ of 12 points for males with right lesions is nearly twice the value of 6.5 points for females. For males, the six-point $P > V$ difference for patients with left damage supports the hypothesis of depressed Verbal IQ for left hemisphere lesions. However, the $P > V$ discrepancy of only 1.6 points for

females with left hemisphere lesions does not support the reversed hypothesis (Kaufman, 1990). This difference across genders for adults with brain lesions may indicate that women have different cerebral organization than men. However, data supporting the reason for the interaction with gender is not definitive (Kaufman, 1990).

Turkheimer and Farace (1992) performed a meta-analysis of 12 different studies which used Wechsler IQ data to examine both male and female patients with right or left hemisphere damage, including a variety of etiologies. The researchers noted a problem in the previous literature to be the use of the difference between the PIQ and VIQ in measuring the effects of lesions. The V – P differences are determined by potentially separate effects of each hemisphere on the IQs. Thus, in this meta-analysis, separate VIQ and PIQ means were reported for men and women with either right or left hemisphere lesions (Turkheimer & Farace, 1992). The results of the repeated-measures analysis revealed that left hemisphere lesions produce substantial and roughly equal VIQ deficits in male and female patients, but lower mean PIQ scores in female than in male patients. Right hemisphere lesions produce PIQ deficits in both genders, but lower mean VIQ scores in female patients. Mean scores from Turkheimer and Farace's (1992) data are presented in Table 4.

The main effect indicated by the data presented in Table 4 is that "female patients are more sensitive to lesions in the hemisphere

Table 3 Effects of unilateral brain damage on WAIS/WAIS-R VIQ–PIQ discrepancies.

Group	Sample size	<i>Mean VIQ minus mean PIQ</i>	
		<i>Left damage</i>	<i>Right damage</i>
Stroke			
Men	124	–10.1	+16.8
Women	81	+0.1	+9.5
Total	248	–6.4	+13.5
Tumor (generalized or posterior)	200	+0.2	+8.4
Frontal lobe	104	–2.2	+2.6
Temporal lobe epilepsy			
Preoperative	101	–3.1	+2.4
Postoperative	101	–6.4	+6.0
Acute lesion	109	–2.4	+14.2
Chronic lesion	131	–2.5	+5.5
Age 20–34	664	–5.0	+6.7
Age 35–54	1245	–3.9	+9.5
Age 55+	168	–2.9	+14.9
Whites	50	–5.2	+15.1
Blacks	50	+5.7	+10.4

Source: Kaufman (1990).

Table 4 Gender differences and brain damage on WAIS/WAIS-R.

	<i>Men</i>	<i>Women</i>
Left damage		
VIQ	91	91
PIQ	95	91
V-P	-4	0
Right damage		
VIQ	104	99
PIQ	90	91
V-P	+ 14	+ 8

Source: Turkheimer and Farace (1992).

Note: V-P = Verbal IQ minus Performance IQ.

Total sample size = 983.

opposite to that thought to be 'odominant' for a function" (Turkheimer & Farace, 1992, p. 499). Although these results are consistent with previously reported greater V-P differences in males, the analysis show that there is also no difference between male and female patients in the effects of left hemisphere lesions on VIQ, or right hemisphere lesions on PIQ. The females demonstrated a pattern of lower mean scores following lesions to hemisphere opposite to the "dominant" hemisphere for each function. This pattern is supportive of a model which asserts that there is a greater degree of bilateral processing in women (Turkheimer & Farace, 1992). This gender difference could be the result of many things including: degree of hemispheric lateralization, differences in problem-solving strategy, or callosal function.

According to Turkheimer, Farace, Yfo, and Bigler (1993), two major findings have been suggested by earlier studies. Individuals with lesions in the left hemisphere have smaller Verbal IQ-Performance IQ differences than subjects with lesions in the right hemisphere, and this difference is greater for males than females. Theories of why gender differences exist can be evaluated through the lesion data. The degree of lateralization in males and females cannot account for gender differences in PIQ and VIQ, because a "statistical model in which the genders have the same degree of lateralization fits the data as well as a model in which the genders are allowed to differ" (Turkheimer et al., 1993, p. 471). There was also not support for the hypothesis that the gender difference results from differences in the within-hemisphere organization of verbal skills.

In a study examining 64 patients through archival data, Turkheimer et al. (1993) did find Inglis and Lawson's (1982) hypothesis that

females use verbal strategies in solving PIQ items to be supported by their data. In females, a single model of lesion effects could account for deficits in VIQ and PIQ, but this was not found for males. The most striking observation made was that females with left-hemisphere lesions had substantial deficits in PIQ related to lesion parameters, but males with left-hemisphere lesions did not (Turkheimer, 1993). Notably, this difference could be present because in the left-hemisphere females may have more non-verbal abilities relevant to PIQ, or females may use more verbal strategies in solving PIQ items. Further research examining problem-solving strategy is necessary to clarify the reason for these gender differences.

(ii) Overview

The WAIS-R has proven itself as a leader in the field of adult assessment. Kaufman (1985) stated, "The WAIS-R is *the* criterion of adult intelligence, and no other instrument is even close." Matarazzo (1985) had an equally favorable review of the WAIS-R, applauding its strong psychometric qualities and clinical usefulness. It has strong reliability and validity for Verbal, Performance, and Full Scale IQs, as did its predecessor, the WAIS. The separate subtests, however, have split-half reliability coefficients that are below 0.75 for six of the 11 tasks at ages 16-17, and for tasks across the age range (Picture Arrangement and Object Assembly) (Kaufman, 1985). The sample selection included apparent systematic bias in the selection of individuals ages 16-19, leading to very questionable teenage norms (Kaufman, 1985). However, the rest of the sample selection was done with precision, leading to an overall well-stratified sample.

Administration is not difficult with the clear and easy to read WAIS-R manual, which provides good scoring directions (Spruill, 1984). The administration and scoring rules of the WAIS-R were made more uniform with the WISC-R rules, which facilitates transfer (Kaufman, 1985). In addition, for the Verbal items with subjective scoring systems, the scoring criteria has been expanded to reduce ambiguity; and to facilitate administration, all words spoken by the examiner are placed on separate lines of the administration manual (Kaufman, 1985).

The WAIS-R does have its limitations; some of which are the nonuniformity of the scaled scores, and the limited floor and ceiling (Spruill, 1984). Individuals who are extremely gifted or severely retarded cannot be assessed adequately with the WAIS-R because the range of possible Full Scale IQ scores is only 45-150. Several

subtests have insufficient floors for adequate assessment of retarded individuals: Picture Arrangement, Arithmetic, Similarities, Picture Completion, and Block Design (Kaufman, 1985). If evaluating an individual who falls at the extreme low end, this is a distinct disadvantage. In addition, even if a subject receives a raw score of zero on a subtest, they can receive one scaled-score point on that subtest.

The WAIS's method of using a reference group (ages 20–34) to determine everyone's scaled scores was retained in the development of the WAIS-R. Kaufman (1985) stated that this method is "indefensible," because use of this single reference group impairs profile interpretation below age 20 and above 34. Profile interpretation is further impaired for individuals aged 16–17 because of low subtest reliability. The WAIS and WAIS-R studies at ages 35–44 cannot be generalized to individuals aged 16–19 because of poor teenage norms, again negatively impacting a clinician's ability to interpret the profile. In the WAIS-R appendix, clinicians must utilize separate scaled-score tables which are arranged by age group. These separate tables invite clerical errors and confusion in case reports (Kaufman, 1985). The WAIS-R manual itself fails to provide appropriate empirical guidelines for profile interpretation, showing a limited awareness of clinicians' practical needs (Kaufman, 1985). However, despite these limitations, the WAIS-R is still one of the most readily chosen instruments in the assessment of intelligence.

4.08.2.1.5 Wechsler Adult Intelligence Scale-Third Edition (WAIS-III)

(i) Description

The newest member of the Wechsler family of tests is the WAIS-III (The Psychological Corporation, 1997). The WAIS-III is an instrument for assessing the intellectual ability of individuals aged 16–89. Like the other Wechsler scales discussed, three IQ scores (Verbal, Performance, and Full Scale) and four factor indices (Verbal Comprehension, Perceptual Organization, Working Memory, and Processing Speed) are derived from the WAIS-III subtests. Each of these scores are standard scores with a mean of 100 and a standard deviation of 15, which are created by comparing an individual's score to scores earned by the normative group of the same age. The Verbal IQ is comprised of six verbal subtest scores (Vocabulary, Similarities, Arithmetic, Digit Span, Information, and Comprehension), plus a new supplementary test to substitute for Digit Span if necessary (Letter–Number Sequencing).

The Performance IQ is comprised of five non-verbal subtests (Picture Completion, Picture Arrangement, Block Design, Matrix Reasoning, and the renamed Digit Symbol-Coding). In addition, two supplemental subtests are provided on the Performance scale: Symbol Search (which may be used to replace Digit Symbol-Coding) and Object Assembly (which is an optional subtest that may be used to replace any performance subtest for individuals younger than 75). In addition to its new name, Digit Symbol-Coding also has two new optional procedures not used in IQ computation, which may be used to help the examiner rule out potential problems. These new procedures include Digit Symbol-Incidental Learning and Digit Symbol-Copy. The Full Scale represents the average of the Verbal and Performance IQs.

New to the WAIS-III are additional factor indices, which can be helpful in further breaking down and understanding an individual's performance. Like the WISC-III, there are four factor indices: Verbal Comprehension, Perceptual Organization, Working Memory, and Processing Speed. The two new subtests on the WAIS-III, Letter–Number Sequencing and Symbol Search, are used in calculation of the Working Memory and Processing Speed Indices, respectively. Table 5 shows which tests comprise each of the IQs and factor indices.

(ii) Standardization properties of the scale

The WAIS-III was standardized by administering the full scale to 2450 adult subjects, selected according to 1995 US Census data. Subjects were stratified according to age, sex, race/ethnicity, geographic region, and education level. Subjects were divided into 13 age groups, which is an improvement over the nine age groups tested in the WAIS-R standardization sample. The number in each WAIS-III standardization age group ranged from 100 to 200, and the age groups spanned from age 16 to 89. Due to the fact that US citizens are living longer, the WAIS-III developers extended norms beyond the highest age group (74) provided in the WAIS-R. In the collection of normative data, an additional 200 African American and Hispanic individuals were also administered the WAIS-III without discontinuation rules. "This over sampling provided a sufficient number of item scores across all items for item bias analyses" (The Psychological Corporation, 1997, p. 20).

Reliability coefficients for the 14 subtests and the Verbal, Performance, and Full Scale IQs were computed using the split-half method (except for Digit Symbol-Coding and Symbol Search). Average reliability, across the 13 age

Table 5 Subtests comprising WAIS-III IQs and Index Scores.

<i>Subtest</i>	<i>IQ scale</i>	<i>Factor index</i>
Vocabulary	VIQ	VCI
Similarities	VIQ	VCI
Information	VIQ	VCI
Comprehension	VIQ	
Arithmetic	VIQ	WMI
Digit Span	VIQ	WMI
Letter-number Sequencing ^a		WMI
Picture Arrangement	PIQ	
Picture Completion	PIQ	POI
Block Design	PIQ	POI
Matrix Reasoning	PIQ	POI
Digit Symbol-coding	PIQ	
Symbol Search ^a		PSI
Object Assembly ^a		PSI

Note. Verbal IQ (VIQ); Performance IQ (PIQ); Verbal Comprehension Index (VCI); Perceptual Organization Index (POI); Working Memory Index (WMI); Processing Speed Index (PSI).

^a The Letter-Number Sequencing, Symbol Search, and Object Assembly subtests can substitute for other subtests under certain circumstances (see The Psychological Corporation, 1997).

groups, is as follows: 0.97 for the Verbal IQ; 0.94 for the Performance IQ; and 0.98 for the Full Scale IQ. The average individual subtests' reliabilities ranged from 0.93 (Vocabulary) to 0.70 (Object Assembly), with a median coefficient of 0.85. Stability coefficients for the three IQs are: 0.96, 0.91, and 0.96 for Verbal, Performance, and Full Scale, respectively. The stability coefficients for individual subtests ranged from an average of 0.94 (Information) to 0.69 (Picture Arrangement) with a median coefficient of 0.815.

The WAIS-III manual (The Psychological Corporation, 1997) reports that numerous factor analytic studies (exploratory and confirmatory) examined the underlying structure of the WAIS-III. There were four basic factors predicted to be underlying the WAIS-III: Verbal Comprehension, Perceptual Organization, Working Memory, and Processing Speed. Overall, results of exploratory and confirmatory factor analysis support the appropriateness of separating the tests of the WAIS into the four factors. The manual reports that the four factor model is a "clearly superior solution to a one-, two-, or three-factor solution and more parsimonious than a five-factor one" (p. 110). Except for the oldest age group, the findings across all ages are similar. However, in the 75–89 year age range, many more subtests loaded on the Processing Speed factor than the Perceptual Organization Factor (i.e., Block Design, Picture Completion, and Picture Arrangement all load on Processing Speed). Only Matrix Analogies had a factor loading above 0.40 on the Perceptual Organization factor for

the oldest age group. From the data presented with the standardization sample, it appears that the WAIS-III is best represented by the four factors that were originally predicted to underlie it.

Across all ages, the Verbal Comprehension factor was defined by loadings ranging from 0.76 to 0.89, including: Information, Vocabulary, Comprehension, and Similarities. Picture Completion, Block Design, Matrix Reasoning, and Picture Arrangement comprised the Perceptual Organization factor with loadings in the 0.47–0.71 range. The third factor is comprised of Digit Span, Arithmetic, and Letter-Number Sequencing with factor loadings of 0.71, 0.51, and 0.62, respectively. Symbol Search and Digit Symbol-Coding are assumed in the Processing Speed factor, with loadings of 0.63 and 0.68, respectively. The Symbol Search subtest requires the examinee to determine whether a pair of target symbols are present in a larger group of shapes within a specified time limit. The addition of the new subtests seems to have strengthened the factor structure, as in the previous version of the WAIS, some of the subtests did not load strongly on any of the factors or loaded similarly across the factors (i.e., Picture Arrangement and Digit Symbol).

(iii) Preliminary research with the WAIS-III

The WAIS-R and the WAIS-III were compared to see how well they were related (The Psychological Corporation, 1997). A sample of 192 individuals with a mean age of 43.5 years (ranging from 16 to 74) were administered the

two tests. The median time between administrations was 4.7 weeks. As would be predicted by work done by Flynn (1984), subjects scored 2.9 points lower on the WAIS-III FSIQ than on the WAIS-R FSIQ. The WAIS-III VIQ and PIQ were 1.2 points and 4.8 points lower than the respective WAIS-R scales. The overall correlations between the WAIS-III and WAIS-R global scales were high. The correlation coefficients for the VIQ, PIQ, and FSIQ were 0.94, 0.86, and 0.93, respectively.

The WAIS-III and WISC-III were also administered to a sample of adolescents to determine how well the two tests correlated (The Psychological Corporation, 1997). The sample consisted of 184 16-year olds who were administered the two tests from 2 to 12 weeks apart (median time = 4.6 weeks). The correlations between the global scales of the two tests were very high, indicating that the two instruments appear to be measuring very similar constructs. The VIQ, PIQ, and FSIQ correlation coefficients were 0.88, 0.78, and 0.88, respectively. The Index scores from the WAIS-III and WISC-III were also compared. The Indices' correlations were 0.87, 0.74, 0.80, and 0.79 for the VCI, POI, WMI, and PSI, respectively. The differences between the mean WISC-III and WAIS-III IQs were all less than one point. The differences between the two tests' mean VCI and POI were also each less than one point. The difference between the WAIS-III and WISC-III mean WMI was 1.7 standard score points. On the PSI, the difference between the means on the two tests was 2.7 points. Thus, overall, the IQ and Indices of the two tests correspond quite highly.

The WAIS-III Technical Manual (The Psychological Corporation, 1997) also presents some studies collected from clinical groups with neurological, psychiatric, and developmental disorders. Reviewed here are a select group of these studies, including those from a sample of patients with mild Alzheimer's disease, a sample of individuals who are mentally retarded, and one from individuals with attention-deficit hyperactivity disorder (ADHD).

Individuals with probable Alzheimer's disease ($N = 35$) were administered both the WAIS-III and the Wechsler Memory Scales-Third Edition (WMS-III). Decrements in cognitive ability and memory were predicted. This sample was reported to have a significantly higher level of education than the normal population, with 48.6% of the sample having completed at least four years of college. The results of this study (The Psychological Corporation, 1997) show that the individuals with probable Alzheimer's disease had lower scores on all IQ scales than the general population. The

mean VIQ, PIQ, and FSIQ scores were 92.2, 81.7, and 86.6, respectively. The PIQ scores tend to be more sensitive to the effects of this neurological condition; thus, the mean VIQ score is predictably higher than the PIQ score. The mean factor index scores demonstrate more differentiation in their cognitive profile. This sample had a mean factor indices of 79.6 (PSI), 84.8 (POI), 87.2 (WMI), and 93.0 (VCI).

A total of 108 individuals diagnosed with mental retardation were administered the WAIS-III. Six of these, 46 were categorized as having mild mental retardation, while the other 62 had moderate mental retardation (The Psychological Corporation, 1997). The results demonstrated deficits across all areas of cognitive functioning, as expected. In the mildly mentally retarded group, mean IQ scores were as follows: 60.1 (VIQ), 64.0 (PIQ), and 58.3 (FSIQ). The subjects with moderate mental retardation exhibited lower scores, earning mean VIQ, PIQ, and FSIQ scores of 54.7, 55.3, and 50.9, respectively. The variability in scores of each of these clinical groups is much smaller than found in the general population. The standard deviations are more than 50% smaller than those found in the general population.

ADHD sufferers was another group studied and reported in the WAIS-III technical manual (The Psychological Corporation, 1997). Traditionally, IQ scores have not been useful in discriminating ADHD from non-ADHD individuals. However, examining subtests patterns on tests of cognitive ability has been more fruitful in discriminating those with ADHD from those without.

The WAIS-III was administered to 30 individuals diagnosed with ADHD (mean age 19.8 years). The mean level of intellectual functioning was found to be in the Average range for this group (mean FSIQ = 103.00). In addition, there was no significant difference found between Verbal and Performance IQs for the group (104.2 and 100.9, respectively). The WAIS-III factor indices were also examined, and the pattern of performance on the indices was found to differ in comparison to the general population. The ADHD sample scored on average 8.3 points lower on the WMI than the VCI. About 30% of the ADHD sample had WMI scores at least 1 SD lower than their VCI scores, whereas 13% of the WAIS-III standardization sample had obtained discrepancies of this magnitude. On average, the ADHD sample scored 7.5 points lower on the PSI than the POI. In the ADHD sample, 26% of the group had PSI scores at least 1 SD lower than their POI scores, but only 14% of the WAIS-III standardization sample had such discrepancies.

(iv) Analyzing the WAIS-III data

The WAIS-III manual provides a general description of how a clinician may begin to interpret the plethora of data obtained in its 14 subtests. However, to obtain the maximum amount of information from the profile, one should utilize an approach to profile interpretation that will group and regroup subtests (Kaufman & Lichtenberger, in press). Similar to the WISC-III interpretation, the WAIS-III may be examined from the global level (IQs) to the individual profile (subtest) level. An organized, systematic approach will be advantageous in obtaining the most accurate picture of the individual.

The WAIS-III record form provides a nice beginning to profile interpretation. However, a structure is needed to work through the large amount of data in a systematic fashion. Kaufman and Lichtenberger, (in press) have developed a series of 10 steps to aid the clinician in interpreting and integrating all of the data obtained from the WAIS-III's three IQs and four factor indices, while not becoming overwhelmed with the multiple scores and difference scores. The 10 steps, a step-by-step approach to WAIS-III profile interpretation, is presented in summary form below. The 10 steps are presented in Table 6. Using these 10 steps can help to organize the information to generate meaningful hypotheses about personal cognitive strengths and weaknesses in preparation for clear presentation in the form of a written report.

(v) Overview

WAIS-III is likely to follow in its footsteps of the WAIS-R, which has proven itself as a leader in the field of adult assessment. The new norms and psychometric improvements of the WAIS-III are much welcomed by the assessment community. The WAIS-III has strong reliability and validity for Verbal, Performance, and Full Scale IQs, as did its predecessor, the WAIS-R. The subtest with the lowest split half reliability has been removed from the computation of the IQs (Object Assembly). However, Picture Arrangement still exhibits split half reliability coefficients below 0.75 at several ages, and it remains part of the Performance IQ. The WAIS-III sample selection was done with precision, leading to an overall well-stratified sample.

Many visual and practical improvements were made in the development of the WAIS-III. Administration is not difficult with the clear and easy to read WAIS-III manual, in addition to the record form with ample space and visual icons (Kaufman & Lichtenberger, in press). The

administration and scoring rules of the WAIS-III were made more uniform throughout the entire test, which reduces chances of examiner error. In addition, the scoring rules are listed right in the administration manual for the Verbal items with subjective scoring systems. This change from the WAIS-R has eased the work of clinicians (no more flipping back and forth during administration).

The WAIS-III attempted to improve its floor and ceiling in comparison to the earlier version. Several step-down items have been added on each subtest for lower functioning individuals. However, like the WAIS-R, individuals who are extremely gifted or severely retarded cannot be adequately assessed with the WAIS-III because the range of possible Full Scale IQ scores is only 45–150. Studies on individuals in the lower extreme range of the WAIS-III are yet to determine whether evaluating an individual who falls at the extreme low end is a distinct disadvantage. As on the WAIS-R, even if a subject receives a raw score of zero on a subtest, they can receive one to five scaled-score points on that subtest. Uniformity is not found across the range of scaled scores for each subtest. On certain subtests subjects may reach a ceiling more quickly than on others. At certain ages, ordinarily the highest scaled score that can be obtained on the subtest is 19; however, on the Arithmetic or Picture Arrangement subtest, 17 is the maximum score (Kaufman & Lichtenberger, in press). Profile analysis is made difficult because of this nonuniformity across subtests, especially for the extremely gifted subjects.

The method of the WAIS-R of using a reference group (ages 20–34) to determine everyone's scaled scores was not retained in the development of the WAIS-III. Kaufman (1985) stated that this WAIS-R method is "indefensible," because use of this single reference group impairs profile interpretation below age 20 and above 34; thus, the WAIS-III change in determining scaled scores is a significant improvement. The process of profile interpretation has been made much less confusing by the removal of the reference group scores. (However, if one wants to calculate the scores by using the 20–34 reference group, this is still possible.) Fewer clerical errors and less confusion in case reports will be present because of these changes (Kaufman & Lichtenberger, in press). The WAIS-III manual and record form themselves provide the beginning to interpretation, with clearly laid out tables to calculate score discrepancies, and so forth. However, to meet clinicians' practical needs for more specific empirical guidelines for profile interpretation in a systematic and step-by-step fashion, other

Table 6 Ten tips for WAIS-III interpretation.*Step 1. Interpret the Full Scale IQ.*

- (i) Report the FSIQ confidence interval (The Psychological Corporation, 1997; Table A.5 p. 197).
- (ii) Report the FSIQ percentile rank (The Psychological Corporation, 1997; Table A.5 p. 197).
- (iii) Report the FSIQ ability level (The Psychological Corporation, 1997, Table 2.3 p. 25).
- (iv) If in STEPS 2–6, it is determined that there is a significant difference between the component parts of the FSIQ (i.e., VIQ & PIQ or VCI & POI), the FSIQ should not be interpreted as a meaningful representation of the individual's overall performance.

Step 2. Determine if the Verbal–Performance IQ discrepancy is statistically significant.

- (i) For all ages VIQ–PIQ difference of 6 points is significant at 0.15 level.
- (ii) For all ages VIQ–PIQ difference of 9 points is significant at 0.05 level.
- (iii) Values for specific ages are presented in Table B.1 (The Psychological Corporation, 1997; p. 205).

*Step 3. Determine if the VIQ and PIQ are interpretable. Four questions to consider about the Verbal and Performance Scales.***Verbal Scale**

- (i) Is the difference between VCI and WMI statistically significant ($p < 0.05$)?
Size needed for difference = 10 + points.
- (ii) Is there abnormal scatter among the VIQ subtests?
Highest of 6 VIQ subtest scaled scores minus lowest = 8 + points.

Performance Scale

- (iii) Is the difference between POI and PSI statistically significant ($p < 0.05$)?
Size needed for difference = 13 + points.
- (iv) Is there abnormal scatter among the PIQ subtests?
Highest of 5 PIQ subtest scaled scores minus lowest = 8 + points.

Step 4. Determine if VIQ–PIQ discrepancy is interpretable or if the VCI and POI should be interpreted instead.

- (i) If all answers in Step 3 are *no*, then the VIQ–PIQ discrepancy is interpretable. Skip Step 5 and go directly to Step 6.
- (ii) If the answer to one or more of the questions in Step 3 is *yes*, then the VIQ–PIQ discrepancy may not be interpretable.
- (iii) If VIQ–PIQ is not interpretable, then examine the VCI–POI discrepancy in Step 5.

Step 5. Determine whether VCI and POI are interpretable and significantly different from one another.

- (i) Is there abnormal scatter among the VCI subtests?
Highest of 3 VCI subtest scaled scores minus lowest = 5 + points.
- (ii) Is there abnormal scatter among the POI subtests?
Highest of 3 POI subtest scaled scores minus lowest = 6 + points.
- (iii) If the answer to either (i) or (ii) is *yes*, then VCI–POI difference may not be interpretable. Otherwise, if both answers are *no*, examine the interpretable VCI–POI difference:
 - (a) Is the difference between VCI and POI statistically significant ($p < 0.05$)?
 - (b) Size needed for difference = 10 + points.

Step 6. Determine if the VIQ–PIQ discrepancy (or the VCI–POI discrepancy) is abnormally large.

- (i) 17 Point difference is abnormally large for the VIQ–PIQ.
- (ii) 19 Point difference is abnormally large for the VCI–POI.
- (iii) Exact point values according to ability level are available in Appendix D (The Psychological Corporation, 1997, pp. 300–309)
- (iv) If the discrepancies are abnormally large, this indicates that they are too big to ignore (see Steps 4 & 5), and they may be interpreted anyway.

Step 7. Determine whether the Working Memory and Processing Speed indices are interpretable.

- (i) Do not interpret WMI if scatter among the 3 subtests is = 6 + points.
- (ii) Do not interpret PSI if difference among 2 subtests is = 4 + points.

Step 8. Interpret the Global Verbal and Nonverbal Dimensions, as well as the small factors, if they were found to be interpretable.

Study the information and procedures presented in (Kaufman & Lichtenberger, in press).

Step 9. Interpret significant strengths and weaknesses in the WAIS-III subtest profile.

- (i) If the VIQ–PIQ discrepancy is less than 17 points, use the individual's mean of all WAIS-III subtests as the person's midpoint.
- (ii) If the VIQ–PIQ discrepancy is 17 or more points, use 2 separate means:
 - (a) Use the individual's mean of all the Verbal subtests as the midpoint for determining strengths and weaknesses on Verbal subtests;

Table 6 (continued)

-
- (b) Use the individual's mean of all the Performance subtests as the midpoint for determining strengths and weaknesses on Performance subtests.
- (iii) Subtract the individual's mean from each of the subtest scaled scores to determine strengths and weaknesses. Round to the nearest whole number.
- (iv) Values are presented in Table B.3 (The Psychological Corporation, 1997, p. 208) for determining if a subtest significantly deviates from the individual's own mean. The following summary information may also be used to determine significance:
- ± 2 points: Vocabulary
 - ± 3 points: Similarities, Arithmetic, Digit Span, Information, Comprehension, Coding, Block Design, Matrix Reasoning,
 - ± 4 points: Letter-Number Sequencing, Picture Completion, Picture Arrangement, Symbol Search
 - ± 5 points: Object Assembly.
-

Step 10. Generate hypotheses about the fluctuations in the WAIS-III subtest profile.

Review the information presented in Kaufman and Lichtenberger (in press) which detail how to reorganize subtest profiles to systematically generate hypotheses about strengths and weaknesses.

sources are available (Kaufman & Lichtenberger, in press). Undoubtedly, like its predecessors, the WAIS-III is likely to become one of the most readily chosen instruments in the assessment of intelligence.

4.08.2.1.6 Kaufman Assessment Battery for Children (K-ABC)

The K-ABC is a battery of tests measuring intelligence and achievement of normal and exceptional children aged 2.5–12.5. It yields four scales: Sequential Processing, Simultaneous Processing, Mental Processing Composite (Sequential and Simultaneous), and Achievement.

The K-ABC is becoming a frequently used test in intelligence and achievement assessment that is used by both clinical and school psychologists (Kamphaus, Beres, Kaufman, & Kaufman, 1995). In a nationwide survey of school psychologists conducted in 1987 by Obringer (1988), respondents were asked to rank the following instruments in order of their usage: Wechsler's scales, the K-ABC, and both the old and new Stanford–Binets. The Wechsler scales earned a mean rank of 2.69, followed closely by the K-ABC with a mean of 2.55, the L-M version of the Binet (1.98), and the Stanford–Binet Fourth Edition (1.26). Bracken (1985) also found similar results of the K-ABC's popularity. Bracken surveyed school psychologists and found that for ages 5–11 years the WISC-R was endorsed by 82%, the K-ABC by 57%, and the Binet IV by 39% of the practitioners. These results suggest that clinicians working with children should have some familiarity with the K-ABC (Kamphaus et al., 1995).

The K-ABC has been the subject of great controversy from the outset, as evident in the strongly pro and con articles written for a special issue of the *Journal of Special Education* devoted to the K-ABC (Miller & Reynolds,

1984). Many of the controversies, especially those regarding the validity of the K-ABC theory, will likely endure unresolved for some time (Kamphaus et al., 1995). Fortunately, the apparent controversy linked to the K-ABC has resulted in numerous research studies and papers that provide more insight into the K-ABC and its strengths and weaknesses.

(i) Theory

The K-ABC intelligence scales are based on a theoretical framework of Sequential and Simultaneous information processing, which relates to how children solve problems rather than what type of problems they must solve (e.g., verbal or nonverbal). In stark contrast is Wechsler's theoretical framework of the assessment of "g," a conception of intelligence as an overall global entity. As a result, Wechsler used the Verbal and Performance scales as a means to an end. That end is the assessment of general intelligence. In comparison, the Kaufmans emphasize the individual importance of the Sequential and Simultaneous scales in interpretation, rather than the overall Mental Processing Composite (MPC) score (Kamphaus et al., 1995).

The Sequential and Simultaneous framework for the K-ABC stems from an updated version of a variety of theories (Kamphaus et al., 1995). The foundation lies in a wealth of research in clinical and experimental neuropsychology and cognitive psychology. The Sequential and Simultaneous theory was primarily developed from two lines of theory: the information processing approach of Luria (e.g., Luria, 1966), and the cerebral specialization work of Sperry (1968, 1974), Bogen (1969), Kinsbourne (1975), and Wada, Clarke, and Hamm (1975).

The neuropsychological processing model, which originated with the neurophysiological

observations of Luria (1966, 1973, 1980) and Sperry (1968), the psychoeducational research of Das (1973; Das et al., 1975; Das, Kirby, & Jarman, 1979; Naglieri & Das, 1988, 1990), and the psychometric research of Kaufman and Kaufman (1983), possesses several strengths relative to previous models in that it (i) provides a unified framework for interpreting a wide range of important individual difference variables; (ii) rests on a well-researched theoretical base in clinical neuropsychology and psychobiology; (iii) presents a processing, rather than a product-oriented, explanation for behavior; and (iv) lends itself readily to remedial strategies based on relatively uncomplicated assessment procedures (Kaufman & Kaufman, 1983; McCallum & Merritt, 1983; Perlman, 1986).

This neuropsychological processing model describes two very distinct types of processes which individuals use to organize and process information received in order to solve problems successfully: successive or sequential, analytic-linear processing vs. holistic/simultaneous processing (Levy & Trevarthen, 1976; Luria, 1966). These processes have been identified by numerous researchers in diverse areas of neuropsychology and cognitive psychology (Perlman, 1986). From Sperry's cerebral specialization perspective, these processes represent the problem-solving strategies of the left hemisphere (analytic/sequential) and the right hemisphere (Gestalt/holistic). From Luria's theoretical approach, successive and simultaneous processes reflect the "coding" processes that characterize "Block 2" functions.

Regardless of the theoretical model, successive processing refers to the processing of information in a sequential, serial order. The essential nature of this mode of processing is that the system is not totally surveyable at any point in time. Simultaneous processing refers to the synthesis of separate elements into groups. The essential nature of this mode of processing is that any portion of the result is, at once, surveyable without dependence on its position in the whole. The model assumes that the two modes of processing information are available to the individual. The selection of either or both modes of processing depends on two conditions: (i) the individual's habitual mode of processing information as determined by social-cultural and genetic factors, and (ii) the demands of the task (Das et al., 1975).

In reference to the K-ABC, Simultaneous processing refers to the mental ability to integrate information all at once to solve a problem correctly. Simultaneous processing frequently involves spatial, analogic, or organizational abilities (Kaufman & Kaufman, 1983; Kamphaus & Reynolds, 1987). There is

often a visual aspect to the problem and visual imagery used to solve it. A prototypical example of a Simultaneous subtest is the Triangles subtest on the K-ABC, which is similar to Wechsler's Block Design. To solve both of these subtests, children must be able to see the whole picture in their mind and then integrate the individual pieces to create the whole.

In comparison, Sequential processing emphasizes the ability to place or arrange stimuli in sequential or serial order. The stimuli are all linearly or temporally related to one another, creating a form of serial interdependence within the stimulus (Kaufman & Kaufman, 1983). The K-ABC subtests assess the child's Sequential processing abilities in a variety of modes. For example, Hand Movements involves visual input and a motor response, Number Recall involves auditory input with a vocal response, and Word Order involves auditory input and visual response. These different modes of input and output allow the examiner to assess the child's sequential abilities in a variety of ways. The Sequential subtests also provide information on the child's short-term memory and attentional abilities.

According to Kamphaus et al. (1995), one of the controversial aspects of the K-ABC was the fact that it took the equivalent of Wechsler's Verbal Scale and redefined it as "achievement." The Kaufmans' analogs of tests such as Information (Faces & Places), Vocabulary (Riddles and Expressive Vocabulary), and Arithmetic (Arithmetic) are included on the K-ABC as achievement tests and viewed as tasks that are united by the demands they place on children to extract and assimilate information from their cultural and school environment. The K-ABC is predicated on the distinction between problem solving and knowledge of facts. The former set of skills are interpreted as intelligence; the latter is defined as achievement. This definition presents a break from other intelligence tests, where a person's acquired factual information and applied skills influence greatly the obtained IQ (Kaufman & Kaufman, 1983).

(ii) Standardization and properties of the scale

Stratification of the K-ABC standardization sample closely matched the 1980 US Census data on the variables of age, gender, geographic region, community size, socioeconomic status (SES), race or ethnic group, and parental occupation and education. Additionally, unlike most other intelligence measures for children, stratification variables also included educational placement of the child (see Table 7).

Table 7 Representation of the K-ABC standardization sample by educational placement.^a

<i>Educational placement</i>	<i>N</i>	<i>%</i>	<i>%</i>
Regular classroom	1862	93.1	91.1
Speech impaired	28	1.4	2.0
Learning disabled	23	1.2	2.3
Mentally retarded	37	1.8	1.7
Emotionally disturbed	5	0.2	0.3
Other ^b	15	0.8	0.7
Gifted and talented	30	1.5	1.9 ^c
Total K-ABC sample	2000	100.0	100.0

^a Data from US Department of Education, National Center for Education Statistics, 1980, Table 2.7, The Condition of Education, Washington, DC, US Government Printing Office. ^b Includes other health impaired, orthopedically handicapped, and hard of hearing. ^c Data from US Office for Civil Rights, 1980, *State, Regional, and National Summaries of Data from the 1978 Child Rights Survey of Elementary and Secondary Schools* (p. 5). Alexandria, VA: Killalea Associates.

Reliability and validity data provide considerable support for the psychometric aspects of the test. A test-retest reliability study was conducted with 246 children after a two- to four-week interval (mean interval = 17 days). The coefficients for the Mental Processing Composite were 0.83 for age two years, six months through four years, eleven months; 0.88 for ages five years through eight years, eleven months; and 0.93 for ages nine years to 12 years, five months. Test-retest reliabilities for the Achievement scale composite for the same age groups were 0.95, 0.95, and 0.97, respectively (Kamphaus et al., 1995). The test-retest reliability research reveals that there is a clear developmental trend, with coefficients for the preschool ages being smaller than those for the school-age range. This trend is consistent with the known variability over time that characterizes preschool children's standardization test performance in general (Kamphaus & Reynolds, 1987). Split-half reliability coefficients for the K-ABC global scales range from 0.86 to 0.93 (mean = 0.90) for preschool children, and from 0.89 to 0.97 (mean = 0.93) for children aged 5–12.5 (Kamphaus et al., 1995).

There has been a considerable amount of research done on the validity of the K-ABC. The *K-ABC interpretive manual* (Kaufman & Kaufman, 1983) includes the results of 43 such studies. Construct validity was established by looking at five separate topics: developmental changes, internal consistency, factor analysis (principal factor, principal components, and confirmatory), convergent and discriminant analysis, and correlations with other tests. Factor analysis of the Mental Processing Scales offered clear empirical support for the existence of two, and only two, factors at each age level, and for the placement of each preschool and

school-age subtest on its respective scale. Analyses of the combined processing and achievement subtests also offered good construct validation of the K-ABC's three-scale structure (Kaufman & Kamphaus, 1984).

Although the K-ABC and the WISC-III differ from one another in a number of ways, there is strong evidence that the two measures correlate substantially (Kamphaus & Reynolds, 1987). In a study of 182 children enrolled in regular classrooms, the Mental Processing Composite (MPC) correlated 0.70 with WISC-R Full Scale IQ (FSIQ), thus, sharing a 49% overlap in variance (Kamphaus et al., 1995; Kaufman & Kaufman, 1983). There have also been numerous correlational studies conducted with handicapped and exceptional populations that may be found in the *Interpretative manual*. The overall correlation between the K-ABC and the WISC-R range from 0.57 to 0.74, indicating that the two tests overlap a good deal, yet also show some independence (Kamphaus et al.).

(iii) Overview

Although the K-ABC has been the subject of past controversy, it appears that it has held its own and is used often by professionals. The K-ABC is well designed with easy to use easels and manuals. The information in the manuals is presented in a straightforward, clear fashion, making use and interpretation of the K-ABC relatively easy (Merz, 1985). There has been a considerable amount of research done on the validity of the K-ABC and the authors have done a thorough job of presenting much of that information in the manual. The reporting of the reliability and validity data in the manual is complete and understandable. However, there is not enough information presented on the

content validity of the test. The various tasks on the subtests on the K-ABC are based on clinical, neuropsychological, and/or other research-based validity; however, a much clearer explanation of the rationale behind some of the novel subtests would have been quite helpful (Merz, 1985).

The K-ABC measures intelligence from a strong theoretical and research basis, evident in the quality of investigation in the amount of research data presented in the manual (Merz, 1985). The K-ABC was designed to measure the intelligence and achievement of children aged 2.5–12.5 and the research done to date suggests that in fact the test does just that. The Nonverbal Scale significantly contributes to the effort to addressing the diverse needs of minority groups and language handicapped children. Overall, it appears that the authors of the K-ABC have met the goals listed in the interpretative manual and that this battery is a valuable assessment tool (Merz, 1985).

Keith and Dunbar (1984) present an alternate means of interpreting the K-ABC, based on exploratory and confirmatory factor analytic data. The two K-ABC Reading subtests are eliminated in their alternate analysis, and factors labeled Verbal Memory, Nonverbal Reasoning, and Verbal Reasoning are presented. For school-aged children whose Achievement Scale splits in half, this model may help interpret their profile. A problem with the Keith and Dunbar labels is that they do not offer evidence to support their Verbal Memory and Nonverbal Reasoning labels. Keith and Dunbar conclude that considerable caution be used when interpreting K-ABC results. In the K-ABC Interpretive Manual (Kaufman & Kaufman, 1983), it is also stressed that a child's profile may need to be approached from an alternative model, if the author's model does not create a good interpretation of the profile.

4.08.2.1.7 Kaufman Adolescent and Adult Intelligence Test

The Kaufman Adolescent and Adult Intelligence Test (KAIT) (Kaufman & Kaufman, 1993) is an individually administered intelligence test for individuals between the ages of 11 and more than 85 years. It provides Fluid, Crystallized, and Composite IQs, each a standard score with a mean of 100 and a standard deviation of 15.

(i) Theory

The Horn–Cattell theory forms the foundation of the KAIT and defines the constructs believed to be measured by the separate IQs;

however, other theories guided the test development process, specifically the construction of the subtests. Tasks were developed from the models of Piaget's formal operations (Inhelder & Piaget, 1958; Piaget, 1972) and Luria's (1973, 1980) planning ability in an attempt to include high-level, decision-making, more developmentally advanced tasks. Luria's notion of planning ability involves decision-making, evaluation of hypotheses, and flexibility, and "represents the highest levels of development of the mammalian brain" (Golden, 1981, p. 285).

Cattell and Horn (Cattell, 1963; Horn & Cattell, 1966, 1967) postulated a structural model that separates fluid from crystallized intelligence. Fluid intelligence traditionally involves relatively culture-fair novel tasks and taps problem solving skills and the ability to learn. Crystallized intelligence refers to acquired skill, knowledge, and judgments which have been taught systematically or learned via acculturation. The latter type of intelligence is influenced highly by formal and informal education and often reflects cultural assimilation. Tasks measuring fluid ability often involve more concentration and problem solving than crystallized tasks which tend to measure retrieval and application of general knowledge.

Piaget's formal operations depicts a hypothetical-deductive abstract reasoning system that has as its featured capabilities the generation and evaluation of hypotheses and the testing of propositions. The prefrontal areas of the brain associated with planning ability mature at about ages 11–12 years (Golden, 1981), the same ages that characterize the onset of formal operational thought (Piaget, 1972). The convergence of the Luria and Piaget theories regarding the ability to deal with abstractions is striking; this convergence provided the rationale for having age 11 as the lower bound of the KAIT, and for attempting to measure decision making and abstract thinking with virtually every task on the KAIT (Kaufman & Kaufman, 1993).

Within the KAIT framework (Kaufman & Kaufman, 1993), crystallized intelligence "measures the acquisition of facts and problem solving ability using stimuli that are dependent on formal schooling, cultural experiences, and verbal conceptual development" (p. 7). Fluid intelligence "measures a person's adaptability and flexibility when faced with new problems, using both verbal and nonverbal stimuli" (Kaufman & Kaufman, 1993, p. 7). It is important to note that this crystallized-fluid construct split is not the same as Wechsler's (1974, 1981, 1991) verbal–nonverbal split. This was documented in the results of a factor analysis done with the WISC-R and the KAIT

that showed the KAIT crystallized subtests loaded highly on the Crystallized/Verbal factor (0.47–0.78), Fluid subtests loaded 0.51–0.88 on the Fluid factor, and Memory for Block Designs also loads 0.41 on the Perceptual Organization factor (Kaufman & Kaufman, 1993; Kaufman, Ishikuma, & Kaufman, 1994). The KAIT Fluid subtests stress reasoning rather than visual-spatial ability, include verbal comprehension or expression as key aspects of some tasks, and minimize the role played by visual-motor speed for correct responding. In addition, the KAIT scales measure what Horn (1989) refers to as broad fluid and broad crystallized abilities, rather than the purer and more specific skill areas that have emerged in Horn's expansion and elaboration of the original Horn-Cattell Gf-Gc theory.

The Core Battery of the KAIT is composed of three Crystallized and three Fluid subtests, and these six subtests are used to compute the IQs. The Expanded Battery also includes two supplementary subtests and two measures of delayed recall that evaluate the individual's ability to retain information that was learned previously in the evaluation during two of the Core subtests. The Core Battery of the KAIT consists of subtests one through six, and subtests one through 10 comprise the Expanded Battery. Each subtest except the supplementary Mental Status task yields age-based scaled scores with a mean of 10 and a standard deviation of three. Sample and teaching items are included for most subtests to ensure that examinees understand what is expected of them for each subtest.

The delayed recall subtests are administered, without prior warning, about 25 and 45 minutes after the administration of the original, related subtests. The two delayed recall subtests provide good measure of an ability that Horn (1985, 1989) calls TSR (Long-Term Storage and Retrieval). TSR "involves the storage of information and the fluency of retrieving it later through association" (Woodcock, 1990, p. 234).

The Mental Status subtest is comprised of 10 simple questions that assess attention and orientation to the world. Most normal adolescents and adults pass at least nine of the 10 items, but the task has special use with retarded and neurologically impaired populations. The Mental Status subtest may be used as a screener to determine if the KAIT can be validly administered to an individual.

(ii) *Standardization and properties of the scale*

The KAIT normative sample, composed of 2000 adolescents and adults between the ages of

11–94 years was stratified on the variables of gender, racial/ethnic group, geographic region, and SES (Kaufman & Kaufman, 1993).

Mean split-half reliability coefficients for the total normative sample were 0.95 for Crystallized IQ, 0.95 for Fluid IQ, and 0.97 for Composite IQ (Kaufman & Kaufman, 1993). Mean test-retest reliability coefficients, based on 153 identified normal individuals in three age groups (11–19, 20–54, 55–85+), retested after a one-month interval, were 0.94 for Crystallized IQ, 0.87 for Fluid IQ, and 0.94 for Composite IQ (Kaufman & Kaufman, 1993). Mean split-half reliabilities of the four Crystallized subtests ranged from 0.89 to 0.92 (median = 0.90). Mean values for the four Fluid subtests ranged from 0.79 to 0.93 (median = 0.88) (Kaufman & Kaufman, 1993). Median test-retest reliabilities for the eight subtests, based on the 153 people indicated previously, ranged from 0.72 to 0.95 (median = 0.78). Rebus Delayed Recall had an average split-half reliability of 0.91 and Auditory Delayed Recall had an average value of 0.71; their respective stability coefficients were 0.80 and 0.63 (Kaufman & Kaufman, 1993).

Factor analysis, both exploratory and confirmatory, gave strong construct validity support for the Fluid and Crystallized Scales, and for the placement of each subtest on its designated scale. Crystallized IQs correlated 0.72 with Fluid IQs for the total standardization sample of 2000 (Kaufman & Kaufman, 1993). Table 8 summarizes the results of correlational studies involving the KAIT and other well known intelligence tests. The values shown in Table 8 support the construct and criterion-related validity of the three KAIT IQs.

Two separate exploratory joint factor analyses were conducted to analyze the KAIT with both the WISC-R and the WAIS-R. Data were obtained from 118 adolescents in the WISC-R sample and 338 adults in the WAIS-R sample. Two-, three-, four-, five-, and six-factor solutions were examined for each analysis using both varimax and oblimin rotations. Three robust factors came out in the three-factor solutions for both the WISC-R and the WAIS-R. Loadings from the first unrotated principal factor (*g* loadings) along with the three-factor solution for the joint analysis of the KAIT and the WISC-R are presented in Table 9. The analogous data for the KAIT and WAIS-R are presented in Table 10.

The results of these joint analyses indicate that the Wechsler subtests and KAIT subtests are about equal as measures of general intelligence. The KAIT subtests have a mean *g* loading of 0.71 and the Wechsler subtests also have a mean *g* loading of 0.71. The most important finding from these analyses is that the

Table 8 Correlations of the three KAIT IQs with standard scores and IQs yielded by other major intelligence tests

<i>Scale</i>		<i>Age</i>	<i>Sample size</i>	<i>KAIT crystallized</i>	<i>KAIT fluid</i>	<i>KAIT composite</i>
WAIS-R IQ	Verbal	16–19	71	0.85	0.74	0.86
		20–34	90	0.78	0.66	0.78
		35–49	108	0.79	0.74	0.85
		50–83	74	0.85	0.70	0.86
	Performance	16–19	71	0.64	0.70	0.72
		20–34	90	0.60	0.74	0.73
		35–49	108	0.57	0.73	0.73
		50–83	74	0.74	0.66	0.77
	Full scale	16–19	71	0.84	0.79	0.88
		20–34	90	0.77	0.76	0.83
		35–49	108	0.74	0.78	0.85
		50–83	74	0.84	0.70	0.85
WISC-R IQ	Verbal	11–16	118	0.79	0.74	0.83
	Performance	11–16	118	0.67	0.67	0.72
	Full scale	11–16	118	0.78	0.75	0.82
K-ABC	Sequential	11–12	124	0.46	0.44	0.50
	Simultaneous	11–12	124	0.53	0.62	0.63
	Mental processing composite	11–12	124	0.57	0.62	0.66
	Achievement	11–12	124	0.81	0.64	0.82
Stanford–Binet-4	Composite intelligence	11–42	79	0.81	0.84	0.87

Source: Kaufman and Kaufman (1993).

KAIT Fluid subtests and Wechsler Performance subtests seem to measure markedly different constructs. The Fluid and Perceptual Organization factors correlate about as highly with each other as they do the Crystallized/Verbal Factor. The differences between Fluid IQ and Perceptual Organization abilities have been studied and discussed by Woodcock (1990). Woodcock presented evidence that Wechsler's Performance IQ primarily measures Horn's Gv or broad visualization, and not fluid intelligence.

The KAIT benefits from an integration of theories that unite developmental (Piaget), neuropsychological (Luria), and experimental-cognitive (Horn–Cattell) models of intellectual functioning. The theories work well together and do not compete with one another. Together, the theories give the KAIT a solid theoretical foundation that facilitates test interpretation across the broad 11–94-year age range on which the battery was normed.

The changes in crystallized and fluid abilities on the KAIT were examined in a study of 1500 adults aged 17–94 (Kaufman & Horn, 1996). The results of this study indicated that fluid reasoning (Gf) declined steadily across adulthood, and this decline accelerated during the

period beginning at about age 55. This finding that fluid ability reaches a peak in development in young adulthood and declines thereafter, at first quite gradually, but more rapidly as age progresses, is in agreement with results from previous research, although a more steep decline in ability was found in the research directed by Horn (1985). Through the 20s, the measure of crystallized knowledge was found to increase, but then showed no increase or decrease until about age 60. Similar findings have been reported in other cross-sectional investigations of fluid and crystallized measures (Kaufman, Kaufman, Chen, & Kaufman, 1996). After age 60, crystallized knowledge was found to decrease as well. Individual differences in education, gender, and ethnicity were found not to alter the basic findings (Kaufman & Horn, 1996).

The KAIT has also been examined with respect to its relationship to adult interests, as demonstrated on the Strong Interest Inventory (SII; Hansen & Campbell, 1985). Kaufman and McLean (1992, November) examined 936 individuals who were administered both the KAIT and the SII. The SII includes six General Occupational Themes (GOTs), including Realistic, Investigative, Artistic, Social,

Table 9 Exploratory joint principal-factor analysis of the KAIT and the WISC-R ($N = 118$).

Subtest	First unrotated factor (g)	Oblimin factor pattern		
		Crystallized/verbal I	Perceptual organization II	Fluid III
<i>KAIT</i>				
Crystallized				
Definitions	82	47		45
Auditory comprehension	69	51		37
Double meanings	79	47		48
Famous faces	65	78		
Fluid				
Rebus learning	69			64
Logical steps	66			60
Mystery codes	78			88
Memory for block designs	74		41	51
<i>WISC-R</i>				
Verbal				
Information	86	62		
Vocabulary	84	75		
Arithmetic	73			59
Comprehension	67	46		
Similarities	79	50	32	
Performance				
Picture completion	67		43	
Picture arrangement	53		36	
Block design	78		79	
Object assembly	73		80	
Coding	55			32

Source: Kaufman and Kaufman (1993).

Note: Decimal points are omitted. Rotated loadings < 0.25 are omitted; those ≥ 0.4 are in bold print. Correlations among factors are as follows: Factors I and II (0.59); Factors I and III (0.69); Factors II and III (0.65).

Enterprising, and Conventional. The findings indicated that there were two General Occupational Themes that produced substantial mean differences between IQ levels. Individuals with higher IQ were more Investigative and more Artistic (Artistic mean score = 49) than those with average IQ (Artistic mean score = 45) or low IQs (Artistic mean score = 42). The authors explain that in light of the Investigative person's interest in science and in solving abstract problems, the relationship of the Investigative theme to IQ level makes sense. The relationship between the Artistic theme and IQ was not hypothesized by the researchers, but was nonetheless significant, even with the effect of Investigative partialled out.

An examination of the KAIT with the Myers-Briggs Type Indicator has been conducted to understand more clearly the commonly accepted relationship between personality style and cognition (Kaufman, McLean, & Lincoln, 1996). The researchers had hypothesized that subjects who favored Intuition and Thinking on the Myers-Briggs would be more intelligent and would also favor fluid over-crystallized intelligence, compared to

those subjects who favored Sensing and Feeling on the Myers-Briggs. Just as hypothesized, individuals classified as Intuitive earned higher KAIT Composite IQs than those classified as Sensing. However, the Fluid IQ was not found to be favored over the Crystallized IQ, as had been predicted (Kaufman et al., 1996). Thus, a modest association is present between personality dimensions and intellectual ability (as evidenced on the Myers-Briggs and KAIT).

4.08.2.1.8 Overview

The KAIT represents a reconceptualization of the measurement of intelligence that is more consistent with current theories of intellectual development (Brown, 1994). The fluid-crystallized dichotomy, the theory underlying the KAIT, is based on the original Horn-Cattell theory of intelligence, thus offering a firm and well researched theoretical framework (Flanagan, Alfonso, & Flanagan, 1994). The fluid-crystallized dichotomy enhances the richness of the clinical interpretations that can be drawn from this instrument (Brown, 1994). The test materials are well constructed and attractive,

Table 10 Exploratory joint principal-factor analysis of the KAIT and the WAIS-R (N = 338).

		Oblimin factor pattern		
Subtest	First unrotated factor (g)	Crystallized/verbal I	Perceptual organization II	Fluid III
KAIT				
Crystallized				
Definitions	78	65		34
Auditory comprehension	73	69		
Double meanings	79	64		47
Famous faces	68	69		
Fluid				
Rebus learning	65			60
Logical steps	66			57
Mystery codes	64		32	56
Memory for block designs	54			62
WAIS-R				
Verbal				
Information	79	89		
Vocabulary	81	91		
Arithmetic	77	47		
Comprehension	79	78		
Similarities	77	59	32	
Digit Span	60	25		29
Performance				
Picture completion	66		49	
Picture arrangement	64	28	57	
Block design	70		76	
Object assembly	64		80	
Coding	50		60	

Source: Kaufman and Kaufman (1993)

Note: Decimal points are omitted. Rotated loadings <0.25 are omitted; those ≥ 0.4 are in bold print. Correlations among factors are as follows: Factors I and II (0.53); Factors II and III (0.53).

and the manual is well organized and helpful (Dumont & Hagberg, 1994; Flanagan et al., 1994). Furthermore, the test materials are easy to use and stimulating to examinees (Flanagan et al.).

"The KAIT has been standardized by state-of-the-art measurement techniques" (Brown, 1994). The psychometric properties of the KAIT regarding standardization and reliability are excellent and the construct validity evidence that is reported in the manual provides a good foundation for its theoretical underpinnings (Flanagan et al., 1994).

The theoretical assumption that formal operations is reached by early adolescence limits that application of the KAIT with certain adolescent and adult populations (Brown, 1994). If an individual has not achieved formal operations, many of the subtests will be too difficult for them and perhaps frustrating and overwhelming. Examiners should be aware of this limitation when working with such individuals in order to maintain rapport. The KAIT can be a useful assessment tool when working with high-functioning, intelligent individuals;

however, it can be difficult to use with borderline individuals and some elderly clients. Elderly clients' scores on some of the subtests may be negatively impacted by poor reading, poor hearing, and poor memory (Dumont & Hagberg, 1994).

Flanagan et al. (1994) report that the inclusion of only three subtests per scale may limit or interfere with the calculation of IQs if a subtest is spoiled. The usefulness of the Expanded Battery and Mental Status subtest of clinical populations is questionable given the reliability and validity data presented in the manual, suggesting that interpretations be made with caution (Flanagan et al.).

Although there clearly are some limitations in the use of the KAIT with some populations, overall, the test appears to be well thought out and validated (Dumont & Hagberg, 1994). The KAIT represents an advancement in the field of intellectual assessment with its ability to measure fluid and crystallized intelligence from a theoretical perspective and, at the same time, maintain a solid psychometric quality (Flanagan et al., 1994).

4.08.2.1.9 *The Stanford–Binet: Fourth edition*

(i) *Theory*

Like its predecessors, the fourth edition is based on the principal of a general ability factor, *g*, rather than on a connection of separate functions. The fourth edition has maintained, albeit to a much lesser degree, its adaptive testing format. No examinee takes all the items on the scale, nor do all examinees of the same chronological age respond to the same tasks. Like its predecessors, the scale provides a continuous appraisal of cognitive development from ages two through young adult.

One of the criticisms in the previous versions is that they tended to underestimate the intelligence of examinees whose strongest abilities did not lie in verbal skills (or overestimate the intelligences of those whose verbal skills excelled). Therefore, consideration when developing the Binet IV was to give equal credence to several areas of cognitive functioning. The authors set out to appraise verbal reasoning, quantitative reasoning, abstract/visual reasoning, and short-term memory (in addition to a composite score representing *g*).

This model is based on a three-level hierarchical model of the structure of cognitive abilities. A general reasoning factor is at the top level (*g*). The next level consists of three broad factors: crystallized abilities, fluid analytic abilities, and short-term memory. The Horn–Cattell theory forms a foundation for the test, with measures of *Gc* being Verbal and Quantitative, and Abstract-Visual being a *Gf* scale. The third level consists of more specific factors, similar to some of Thurstone's eight primary mental factors: verbal reasoning, quantitative reasoning, and also abstract/visual reasoning. The selection of these four areas of cognitive abilities came from the authors' research and clinical experience of the kinds of cognitive abilities that correlate with school progress.

The Binet IV contains previous tasks, combining old with new items, and some completely new tasks. In general, test items were accepted if (i) they proved to be acceptable measurements of the construct, (ii) they could be reliably administered and scored, (iii) they were relatively free of ethnic and/or gender bias, and (iv) they functioned adequately over a wide range of age groups.

(ii) *Standardization and properties of the scale*

Standardization procedures followed 1980 US Census Data. There appears to be an accurate sample representation from geographic region, size of community, race/ethnic group, and gender. The standardization falls

short, however, in terms of age, parental occupation, and parental education. The total sample size was large (5013), with age representation extending from two years to 23 years, 11 months. The concentration of the sample is on children aged 4–9 years old (41%). Not only were adults 24 years and older not represented, but also representation beyond age 17 years, 11 months was negligible (4%).

In order to assess characteristics of SES, information regarding parental occupation and parental education was obtained. A review of Table 11 demonstrates that children whose parents came from managerial/professional occupations and or who were college graduates and beyond were grossly over-represented in the sample. In other words, the norms are based on a large percentage of individuals from upper-socioeconomic classes. In order to adjust for this discrepancy, a weighting procedure was applied, which makes the norming sample suspect. Unquestionably, SES has been shown time and again to be the single most important stratification variable regarding its relationship to IQ (Kaufman, 1990, Chapter 6; Kaufman & Doppelt, 1976).

Internal consistency estimates for the Stanford–Binet IV Composite Scale are excellent, ranging from 0.95 to 0.99 across the age groups (median = 0.97) (Sattler, 1988). The internal reliabilities are also high for the Verbal Reasoning, Abstract/Visual Reasoning, Quantitative Reasoning, and Short-term Memory Area scores (typically in the upper 0.80s–0.90s). Subtest reliabilities are also good, with the exception of Memory for Objects which had a median of 0.73 (Thorndike et al., 1986b). Test–retest reliability estimates are also good for preschool (Composite coefficient = 0.91) and elementary school aged (Composite coefficient = 0.90) samples (Thorndike, Hagen, & Sattler, 1986a). From an internal reliability perspective, this measure is generally good.

Construct validity for *g* and for the four factors was studied using a variant of confirmatory factor analysis. The subtests had impressive high to substantial loadings on *g* (0.51–0.79). Unfortunately, the four factors were given weak support by the confirmatory procedure. Additionally, exploratory factor analysis gave even less justification for the four Binet Scales; only one or two factors were identified by Reynolds, Kamphaus, and Rosenthal (1988) for 16 of the 17 age groups studied. Clearly, the factor analytic structure does not conform to the theoretical framework used to construct the test. Therefore, once again the composite score is left as the only clearly valid representation of a child's cognitive abilities.

Table 11 Representation of the Stanford–Binet fourth edition.

	Sample percent	US population percent
<i>By parental occupation</i>		
Managerial/professional	45.9	21.8
Technical sales	26.2	29.7
Service occupations	9.7	13.1
Farming/forestry	3.2	2.9
Precision production	6.7	13.0
Operators, fabricators, other	8.3	9.5
Total	100.0	100.0
<i>By parental education</i>		
College graduate or beyond	43.7	19.0
1–3 years of college	18.2	15.3
High school graduate	27.5	36.5
Less than high school	10.6	29.2
Total	100.0	100.0

Correlational studies, using nonexceptional children, between the Binet IV and the Stanford–Binet (Form L–M), WISC–R, WAIS–R, WPPSI, K–ABC have ranged from 0.80 to 0.91 (comparing full-scale composites). Correlational studies using exceptional children (gifted, learning impaired, mentally retarded) produced generally lower correlations, probably because of restricted variability in the test scores. For example, for gifted students the mean composite score on the Binet IV correlated 0.69 with the WISC–R Full Scale IQ. These data and data from similar validity investigations are presented more extensively in the *Technical manual* for the Binet IV (Thorndike et al., 1986a). Despite the presentation of ample evidence of concurrent validity, the substantial problems with construct validity, the data collection method, and other difficulties with the Binet IV have led at least one reviewer to recommend that the battery be laid to rest (Reynolds, 1987): “To the S–B IV, *Requiescat in pace*” (p.141).

(iii) Overview

The Binet IV was developed in an attempt to increase the popularity of the test as well as address some of the negative reviews that had plagued the previous edition. The test authors attempted to make the fourth edition significantly different from the previous L–M edition; however, it appears that this goal has achieved only limited success. Canter (1990) describes the “rebirth” of the Binet as giving way to “confusion and even dismay as the primary consumers of intelligence tests learned that the new edition offered a more complicated route to the same destination.” Another reviewer de-

scribes the Binet IV as “in most respects, a completely new version of a very old test” (Spruill, 1987). This author also questioned whether or not the weighting procedure that was used to correct for sample bias was adequate (Spruill, 1987) that was not outweighed by the large size of the standardization sample. Finally, it is not clear why a test described as for individuals aged two to adult does not include persons over the age of 23 in the standardization sample.

Although there appears to be a number of difficulties with the Binet IV, the test is still used and it is not without its strengths. The administration of some of the subtests allow the examiner flexibility, and young children seem to find the items challenging and fun. The scale has excellent internal reliability and provides a flexible administration format. Despite its shortcomings, Binet IV continues to be a very good assessment of cognitive skills related to academic progress (Spruill, 1987). It also includes several excellent, well-constructed tasks that offer valuable information when they are administered in addition to the Wechsler scales (Kaufman, 1990, 1994b).

4.08.2.1.10 Woodcock–Johnson Psycho-Educational Battery–Revised: tests of cognitive ability (WJ–R)

The WJ–R is one of the most comprehensive test batteries available for the clinical assessment of children and adolescents (Kamphaus, 1993). The WJ–R is a battery of tests for individuals from age 2 to 90+, and is composed of two sections, Cognitive and Achievement. The focus of this discussion is the Cognitive portion of the WJ–R battery.

(i) Theory

The WJ-R Cognitive battery is based on Horn's (1985, 1989) expansion of the Fluid/Crystallized model of intelligence (Kamphaus, 1993; Kaufman, 1990). The standard and supplemental subtests of the WJ-R are aligned with eight of the cognitive abilities isolated by Horn (1985, 1989) (Kamphaus, 1993; Kaufman 1990). The cognitive battery measures seven Horn abilities: Long-term Retrieval, Short-term Memory, Processing Speed, Auditory Processing, Visual Processing, Comprehension-knowledge, and Fluid Reasoning. An eighth ability, Quantitative Ability, is measured by several Achievement subtests on the WJ-R.

The four subtests that measure Long-term Retrieval (Memory for Names, Visual-Auditory Learning, Delayed Recall/Memory for Names, Delayed Recall/Visual-Auditory Learning), require the subject to retrieve information stored minutes or a couple of days earlier. In contrast, the subtests that measure Short-term Memory (Memory for Sentences, Memory for Words, Numbers Reversed) require the subject to store information and retrieve it immediately or within a few seconds. The two Processing Speed subtests (Visual Matching, Cross Out) assess the subject's ability to work quickly, particularly under pressure to maintain focused attention.

Within the Auditory Processing domain, three subtests (Incomplete Words, Sound Blending, Sound Patterns) assess the subject's ability to perceive fluently patterns among auditory stimuli. The three Visual Processing subtests (Visual Closure, Picture Recognition, Spatial Relations) assess the subject's ability to manipulate fluently stimuli that are within the visual domain.

Picture Vocabulary, Oral Vocabulary, Listening Comprehension, and Verbal Analogies are the four subtests that are linked to the Comprehension-knowledge factor, also known as crystallized intelligence within Horn's theoretical model. These subtests require the subject to demonstrate the breadth and depth of their knowledge of a culture. Analysis-synthesis, Concept Formation, Spatial Relations, and Verbal Analogies (which also loads on the Comprehension-knowledge factor) assess the subject's Fluid Reasoning, or "new" problem-solving ability. Finally, from the Achievement portion of the WJ-R, both the Calculation and Applied Problems subtests assess the individual's Quantitative Ability.

The cognitive battery consists of 21 subtests, seven of which comprise the standard battery; the remaining 14 are part of the supplemental battery (one per ability as described by Horn).

There are two composite scores, Broad Cognitive Ability and Early Development (for pre-schoolers), which are both comparable to an overall IQ. The individual subtest scores as well as the composite scores have a mean of 100 and a standard deviation of 15.

Computer software is available for scoring the WJ-R and is essential if all of the information is to be obtained that the WJ-R is capable of providing. The WJ-R provides the examiner with percentile ranks, grade-based scores, age-based scores, and the Relative Mastery Index (RMI). The RMI is a unique kind of ratio, with the second part of the ratio set at a value of 90. The denominator of the ratio means that children in the norm sample can perform the intellectual task with 90% accuracy. The numerator of the ratio refers to that child or adolescent's proficiency on that subtest (Kamphaus, 1993). For example, if a child obtains an RMI of 60/90, it would mean that the child's proficiency on the subtest is at a 60% level whereas the typical child of that age (or grade) mastered the material at a 90% level of accuracy.

The entire battery is quite lengthy and therefore can be timely to administer. The seven-subtest Standard Battery takes approximately 40 minutes to administer; however, all the clinician will obtain from it is, essentially, a measure of *g*. In order to obtain all of the information that the WJ-R is capable of providing, a clinician should administer most of the subtests in both the Cognitive and Achievement batteries. Administration of a thorough cognitive and achievement assessment using the WJ-R would take approximately 3.5–5 hours depending on the subject's age, abilities, and speed. However, individual subtests may be administered to test specific hypothesis without administering the entire battery. The WJ-R tests also provides measures of differential scholastic aptitudes including reading, mathematics, written language, and knowledge. An aptitude-achievement comparison may be made if the WJ-R Tests of Achievement are given in addition. Such a discrepancy reflects the amount of disparity between certain intellectual capabilities of an individual and their actual academic performance.

Evidence has been presented that supports the use of the WJ-R, standard cognitive and achievement tests in the identification and classification of school-aged children as gifted, learning-disabled, and mentally retarded (Evans, Carlsen, & McGrew, 1993). Significant group differences were found on mean scores of all WJ-R standard cognitive and achievement clusters. Together, the WJ-R cognitive and achievement demonstrated the ability to predict

group membership (gifted, L.D., or M.R.). This was shown in an overall classification agreement of 93.5% for ages 8–10 and 84.3% for ages 16–18. These levels of classification agreement support the use of the WJ-R batteries in the identification of exceptional students (Evans et al., 1993). It should be noted, however, that clinicians must supplement the process of assessment and diagnosis by including other factors beyond statistical classification, such as social-emotional considerations, medical conditions, vision and hearing measures, and other environmental considerations, to make the determination of classification of the aforementioned groups.

(ii) *Standardization and properties of the scale*

The WJ-R was normed on a reasonably representative sample of 6359 individuals selected to provide a cross-section of the US population aged 2–90+ (Woodcock & Mather, 1989). The sample included 705 preschool children, 3245 students in grades K–12, 916 college/university students, and 1493 individuals aged 14–90+ who were not enrolled in school. Stratification variables included gender, geographic region, community size, and race. However, Kaufman (1990) reports that, although representation on important background variables was adequate, it was necessary to use a weighting procedure to adjust the data that was collected so it would match US population statistics.

The internal consistency estimates for the standard and supplemental battery subtests are good, with median scores from ages 2 to 79 ranging from 0.69 to 0.93. The Broad Cognitive Ability composite score for the seven standard battery subtest yields a median internal consistency coefficient of 0.94 and the Broad Cognitive Ability Early Development Scale yields a coefficient of 0.96 at ages two and four (Kamphaus, 1993).

The validity of the WJ-R has been called into question when used with a learning disabilities population (Hoy et al., 1993). This is because the mean broad cognitive ability scores from the WJ-R Tests of Cognitive Ability have been found to be one standard deviation lower than mean full scale scores on the WISC-R. However, in an adult sample, significant and consistently high correlations were found between the WAIS-R and the WJ scores (Siehen, 1985). Hoy et al. studied 27 male and 20 university students with a previously diagnosed learning disability, as well as 47 learning disabled individuals from a rehabilitation clinic. The results indicated a low positive correlation between the WJ-R broad cognitive score and the

WAIS-R Verbal Scale scores ($r = 0.44$) for the rehabilitation subjects, but a more moderate positive correlation was found for the university subjects ($r = 0.73$). The WAIS-R Full Scale IQ also correlated moderately with the Broad Cognitive in university subjects ($r = 0.72$), and had a lower correlation with rehabilitation subjects ($r = 0.58$). The correlation between WJ-R Broad Cognitive Scores and the Performance Scale IQ was low for both the university and rehabilitation subjects ($r = 0.40$). The low correlation between the PIQ and Broad Cognitive suggests that the two instruments are providing different information, and are therefore both potentially useful.

(iii) *Overview*

The WJ-R Cognitive battery was developed based on Horn's expansion of the Cattell–Horn Fluid–Crystallized model of intelligence. This theoretical rationale allows for further empirical analysis of both the WJ-R and the theory (Webster, 1994). The standardization of the battery appears to be sound and the various age groups are represented adequately.

The Cognitive battery is quite thorough and, when administered in its entirety, can provide the examiner with a wealth of information about an individual's intellectual functioning and abilities. The test materials and manuals are easy to use and well designed. The administration is fairly simple; however, scoring the test, especially when the Achievement battery is administered as well, can be quite a lengthy and, initially, a difficult process. The scoring can be done by hand but is done more efficiently with the computer scoring program. The computer scoring program is easy to use and provides the examiner with the individual's raw scores, standard scores, percentile ranks, and age and grade equivalents for each subtest (Webster, 1994).

The WJ-R Cognitive battery is a well standardized test developed on a theory of intelligence. However, the test is not without shortcomings. Webster (1994) raises issues with the specific psychometric procedures used in developing test items. Data are lacking that show the efficacy of the WJ-R to predict, from a time based perspective, actual functional levels of academic achievement and to identify children at risk for failure early in the educational process (Webster, 1994). Kaufman (1990) points to another shortcoming with the small number of tests that comprise each scale. The Standard Scale measures each of the seven scales with one subtest apiece.

The *Woodcock–Johnson Psycho-Educational Battery: Revised examiner's manual* reports that

"Items included in the various tests were selected using item validity studies as well as expert opinion" (Woodcock & Mather, 1989, p. 7). Kamphaus (1993) states that the manual should have included more information on the results of the experts' judgments or some information on the methods and results of the studies that were used to assess validity.

It is clear that the WJ-R Cognitive battery is quite comprehensive, providing the clinician with a wealth of information. The standardization sample is large, the factor loadings reveal generally strong factor analytic support for the construct validity for the battery for adolescents and adults, and the reliability coefficients are excellent (Kaufman, 1990).

4.08.2.1.11 Detroit Tests of Learning Aptitude (DTLA-3)

DTLA-3 was developed by Hammill (1991) and was designed to measure different, but interrelated, mental abilities for individuals aged six through 17 years, 11 months. It is a battery of 11 subtests and yields 16 composites that measure both general intelligence and discrete ability areas. Hammill and Bryant (1991) report that the DTLA-3 was influenced greatly by Spearman's two-factor theory (1927). This theory of "aptitude" consisted of a general factor *g* that is present in all intellectual pursuits, and specific factors that vary from task to task (McGhee, 1993).

The 11 subtests are used to form the 16 composite scores. The subtests are grouped into different combinations according to various hypothetical constructs that exist in theories of intelligence and information processing. In general, the composite scores estimate general mental ability; however, they all do so in a somewhat different manner. The General Mental Ability Composite is formed by combining the standard scores of all 11 subtests, and, thus, has been referred to as the best estimate of *g*. The Optimal Level Composite is composed of the four largest standard scores that the individual earns. This individualized score is often referred to as the best estimate of a person's overall "potential." The Domain Composites may be divided into three areas: Linguistic, Attentional, and Motoric. Furthermore, there is a Verbal and Nonverbal Composite in the Linguistic domain, an Attention-enhanced and Attention-reduced Composite in the Attentional Domain, and a Motor Enhanced and a Motor-reduced composite in the Motoric Domain. Finally, there are the Theoretical Composites of the DTLA-3 on which the battery's subtests are constructed. The major theories that the subtests were

developed from include Horn and Cattell's (1966) fluid and crystallized intelligences, Das' (1973) simultaneous and successive processes, Jensen's (1980) associative and cognitive levels, and Wechsler's (1974, 1981, 1989) verbal and performance scales.

The DTLA-3 yields five types of scores: raw scores, subtest standard scores, composite quotients, percentiles, and age equivalents. Standard scores for the individual subtests have a mean of 10 and a standard deviation of three and the Composite Quotients have a mean of 100 and a standard deviation of 15.

The individual subtest reliabilities range from 0.77 to 0.94, with a median of 0.87, and the averaged alphas for the composites ranged from 0.89 to 0.96, with a median of 0.94. To assess the DTLA-3's stability over time, the test-retest method was used with a sample of 34 children residing in Austin, Texas. The children, aged six through 16, were tested twice, with a two-week period between testings (Hammill, 1991). The results of this test-retest analysis indicate that individual subtest reliabilities range from 0.75 to 0.96, with a median of 0.84, and Composite reliabilities range from 0.81 to 0.96, with a median of 0.90.

(i) Overview

The DTLA-3 was designed to measure both general intelligence and discrete abilities for children aged six through 17 years, 11 months. The DTLA-3 is not grounded in one specific theory but rather can be linked to a number of different theorists and their views on intelligence and achievement. This "eclectic" theorizing has resulted in the DTLA-3's numerous subtests, composites and various combinations of the two that yield potentially important information about an individual's abilities.

Reliability and validity studies are encouraging but are based on specific and limited samples (VanLeirsburg, 1994). Additional research in this area would be beneficial. Furthermore, test-retest reliability data were collapsed across age levels, which makes it impossible to determine the stability of scores of the various age levels (Schmidt, 1994). The standardization sample was representative of the US population but more information on socioeconomic level is needed (Schmidt, 1994). Also, there is no normative data reported for subjects with handicapping conditions and sample stratification for age was not equalized (VanLeirsburg, 1994).

The testing manual suggests that individual testing time may vary but that on average it takes 50 minutes to two hours to administer. Scoring and interpretation of the results is easy,

yet it can be quite time consuming without the aid of the computer program (VanLeirsburg, 1994). Despite apparent shortcomings, the DTLA-3 may be useful for eligibility or placement purposes as well as a useful research tool (Schmidt, 1994).

4.08.2.1.12 Differential Abilities Scales (DAS)

The DAS was developed by Elliott (1990) and is an individually administered battery of 17 cognitive and achievement tests for use with individuals aged 2.5 through 17 years. The DAS Cognitive Battery has a preschool level and a school-age level. The school-age level includes reading, mathematics, and spelling achievement tests that are referred to as "screeners." The same sample of subjects was used to develop the norms for the Cognitive and Achievement Batteries; therefore, intra- and intercomparisons of the two domains are possible.

The DAS is not based on a specific theory of intelligence. Instead, the test's structure is based on tradition and statistical analysis. Nonetheless, the test is not theory free, and, in fact, is based in part on *g* and the view of intelligence as hierarchical in nature (McGhee, 1993). Elliott (1990) described his approach to the development of the DAS as "eclectic" and cited researchers such as Cattell, Horn, Das, Jensen, Thurstone, Vernon, and Spearman. Indeed, there are some clear-cut relationships between several DAS scales and theoretical constructs. For example, Horn's (1985, 1989) concepts of fluid and crystallized intelligence are measured quite well by the Nonverbal Reasoning and Verbal Ability scales, respectively. Elliott endorses Thurstone's ideas that the emphasis on intellectual assessment should be on the assessment and interpretation of distinct abilities (Kamphaus, 1993). He also stresses that in the assessment of children with learning and developmental disabilities, clinicians need more fine detail than is provided by a global IQ score (Elliott, 1997). Therefore, subtests were constructed to emphasize their unique variance which should translate into unique abilities. Although it was expected that meaningful composites would be derived from the subtests, the primary focus in test development was at the subtest level. One of the main distinctions between the DAS and other batteries is its emphasis on the subtest level.

Elliott (1997) noted how in psychology there has long been a link between cognitive abilities and neurological structure. The DAS uses the link between the factor structure of abilities and the neurological evidence of the nature of the structures. For example, the DAS has two major ability clusters which reflect two major

systems for receiving, perceiving, remembering, and processing information in the visual and auditory modalities. The systems are represented by Verbal and Visualization/Spatial factors. There is strong neuropsychological evidence for the existence of these systems (Elliott, 1997), which tend to be specific in the left and right cerebral hemispheres, respectively. The DAS specifically measures each of these factors by the Verbal cluster and the Spatial Cluster.

Normally, the auditory and visual systems do not operate completely independently. There is interaction of the systems. The integrative system is represented factorally by the fluid reasoning factor in the Cattell-Horn theory. Analysis of both verbal and visual information is usually required in measures of fluid reasoning. The neuropsychological function is an integrative system of the frontal lobe, which is critical to complex mental functioning (Luria, 1973). The DAS measures fluid ability by the Nonverbal Reasoning cluster, which requires integrated analysis and transformation of both visual and verbal information.

Elliott (1997) notes that there is much evidence from cognitive psychology that indicates that verbal and visual short-term memory systems are quite distinct. However, some cognitive tests (i.e., Stanford-Binet IV) represent memory with a single factor. The DAS, on the other hand, represents visual and auditory short-term memory with distinct measures, rather than with one unitary task. The DAS also provides a measure of intermediate-term memory (in the Horn-Cattell model), which is usually measured by tasks that have both verbal and visual components. The DAS has a measure in which pictures are presented, but they have to be recalled verbally (Recall of Objects). This visual-verbal short-term memory measures another distinct information processing system (Elliott, 1997).

The cognitive portion of the DAS consists of "core" and "diagnostic" subtests designed to assess intelligence at the preschool level and the school-age level. The core subtests measure complex processing and conceptual ability, which is strongly *g*-related. The diagnostic subtests measure less cognitively complex functions, such as short-term memory and processing speed, thereby having less of a *g*-saturation (Elliott, 1997). The achievement portion measures skills in the areas of word reading, spelling, and basic number skills. The core subtests are averaged to obtain the General Conceptual Ability (GCA) score and, depending on the age of the individual, additional composite scores are calculated which are referred to as Cluster scores.

The individual Cognitive subtests have a mean of 50 and a standard deviation of 10. The GCA scores, Cluster scores, and Achievement scores, have a mean of 100 and a standard deviation of 15. Percentile ranks, age equivalents, and score comparisons are also available in the examiner's manual. Score comparisons provide a profile analysis and allow the examiner to ascertain information regarding aptitude-achievement discrepancies.

Interpretation of the DAS subtests and composites is facilitated by the framework provided in the *Handbook* (Elliott, 1990). Elliott (1990) notes one positive aspect of interpreting the DAS is that the "design of scoring procedures on the Record Form enables statistically significant high and low scores to be identified immediately" (p. 37). Significant discrepancies between subtests, cluster scores, and ability and achievement can be obtained immediately. Like Kaufman's (1994b) approach to interpreting the WISC-III, an ipsative approach is used to examine differences between subtests, requiring the examiner to compare the child's mean score on core subtests to his or her individual subtest scores. Discrepancy between ability and achievement is analyzed by examining the GCA (or Special Nonverbal Composite) and each of the achievement tests.

(i) *Standardization and properties of the scale*

Elliott (1997) noted that exceptionally careful and effective standardization and data-analytic procedures were used in the development of the DAS. The DAS was standardized on 3475 children tested between 1987 and 1989. The normative sample includes 200 cases for each age level between the ages of five and 17. The younger part of the sample consisted of 350 children between the ages of 2.5 year, and four years, 11 months. Exceptional children were also included in the standardization sample. Gender, race, geographic region, community size, and enrollment (for ages 2–5 through 5–11) in an educational program were controlled. SES was estimated using the average education level of the parent or parents living with the child (Kamphaus, 1993).

Over and above the requirements of the norm sample, an additional 600 cases of Black and Hispanic children were collected. The reason that this oversample was collected was to perform statistical analysis for item bias and prediction bias. The test developers wanted to ensure that the rules for scoring would be sensitive to minority children's responses (Elliott, 1997). Only a small number of items were deleted due to item bias, and there was "no evidence" that the DAS is biased against either

Blacks or Hispanics, compared with Whites (Elliott, 1990).

The DAS has a median reliability estimate of 0.95 for the GCA. Internal consistency reliability estimates for the cluster scores range from 0.83 for Nonverbal Reasoning at age five to 0.94 for Spatial at several ages (Kamphaus, 1993). The test-retest reliability coefficients for the preschool composite scores are 0.84 for Verbal Ability and 0.79 for Nonverbal Ability. The individual subtests' reliabilities vary with an average coefficient of 0.78.

Correlational research has shown good evidence of concurrent validity for the DAS (Kamphaus, 1993). With a sample of 27 children aged 7–14, the WISC-III Full Scale IQ correlated very highly with the DAS GCA score (0.92), and the WISC-III Verbal IQ score correlated highly with the DAS Verbal Ability score (0.87). The WISC-III Performance IQ correlated 0.78 with Nonverbal Reasoning and 0.82 with Spatial Ability. Additionally, the DAS Speed of Information Processing subtest score correlated 0.67 with the WISC-III Processing Speed Index score. The Binet IV Composite IQ correlated 0.88 with the DAS GCA for nine- and 10-year-olds and 0.85 with the DAS GCA for a sample of gifted children. The K-ABC Mental Processing Composite correlated 0.75 with the DAS GCA for 5–7-year-olds (Kamphaus, 1993).

Elliott (1997) presents three validity studies not published in the DAS manual. One of these studies included a confirmatory factor analysis of the DAS by Keith (1990), which concluded that "the constructs measured by the DAS are remarkably consistent across overlapping age levels of the test" (Elliott, 1997, p. 20). Elliott also discusses a joint factor analysis of the DAS and WISC-R. In a reanalysis of data, Elliott (in press) reported five factors emerging: crystallized intelligence (including DAS verbal and WISC-R verbal subtests), spatial or broad visualization (including DAS spatial and four of the five major WISC-R Performance subtests), nonverbal reasoning or fluid intelligence (defined only by DAS Nonverbal Reasoning subtests), auditory short-term memory, and speed of processing.

(ii) *Overview*

In general, the professional reviews of the DAS seem to be quite positive. Sandoval (1992) reports that the DAS is one of the least biased tests available. The test appears to be a relatively culture fair measure; however, its use with linguistically different children needs to be explored further (Sandoval, 1992). In examining different cultural group's performance on the

DAS, the group differences found typically to be present on traditional IQ tests are also found on the DAS. For example, African-American and Hispanic children score between half and two-thirds of a standard deviation below White children, and Asian children score above White children on all but verbal areas of the test. Caution is necessary when assessing Hispanic children because the DAS overpredicts achievement for this group based upon group achievement results (Bain, 1991). The author of the DAS suggests that children who are not proficient in English be given the nonverbal tests on the Special Nonverbal scale in the primary language. However, this can be problematic as the test developers did not provide directions in other common languages, such as Spanish. In addition, the utility of the English norms for assessing a child who is administered the test in Spanish or another non-English language has not been explored.

The DAS manual has recommendations for administering the test to deaf or limited-English-proficient children, however, these recommendations are lacking in a couple of areas. Braden (1992) notes, "The recommendations that age equivalents be used to represent the performance of retarded person is common, but it is potentially misleading" (p. 93). Another problem with the recommendations is that no mention is made of the use of interpreters for hearing-impaired children and nonverbal children, which may have a detrimental effect on deaf children's test scores.

According to Braden (1992), the Technical Manual includes extensive research data which suggest that the DAS is a psychometric improvement over existing techniques for measuring intellectual abilities and for determining intracognitive and aptitude-achievement discrepancies. The GCA of the DAS is largely independent of tasks known to be difficult for learning disabled children and is able to assist in the identification of learning disabilities or processing deficits.

The DAS can be a useful tool in assessing intelligence and achievement in both children and adolescents. However, there are a few characteristics of the DAS which do not promote ease of administration, especially for novices (Braden, 1992). These difficulties include having to apply two rules for subtest discontinuation, and having to convert raw scores to ability scores prior to obtaining subtest scaled scores. The DAS examiner's manual provides interpretive information and a framework for interpretation for the composite scores and subtests. The level and/or depth of information that the interpretative portion of the manual provides is quite thorough and is

easy to use, making interpretation of the profiles, and individual and composite scores much easier.

4.08.2.1.13 Cognitive Assessment System (CAS)

(i) Theory

The Das–Naglieri Cognitive Assessment System (Naglieri & Das, 1996) was developed according to the Planning, Attention, Simultaneous, and Successive (PASS) theory of intelligence. The subtests are organized into four scales designed to provide an effective measure of the PASS cognitive processes. Planning subtests require the child to devise, select, and use efficient plans of action to solve the test problems, regulate the effectiveness of the plans, and self-correct when necessary. Attention tests require the child to selectively attend to a particular stimulus and inhibit attending to distracting stimuli. Simultaneous processing tests require the child to integrate stimuli into groups to form an interrelated whole and Successive processing tests require the child to integrate stimuli in their specific serial order or appreciate the linearity of stimuli with little opportunity for interrelating the parts.

The CAS yields Planning, Attention, Simultaneous, Successive, and Full Scale normalized standard scores (mean of 100 and standard deviation of 15). The Planning scale's subtests include Matching Numbers, Planned Codes, Planned Connections, and Planned Search; the Attention scale subtests include Number Detection, Receptive Attention, and Expressive Attention; the Simultaneous Scale subtests are Nonverbal Matrices, Verbal–Spatial Relations, and Figure Memory; and the Successive Scale subtests are Word Series, Sentence Repetition, Sentence Questions, and Successive Speech Rate. All subtests are set at a normative mean of 10 and SD of three.

(ii) Standardization and properties of the scale

The CAS was standardized on 2200 children ranging in age from five through 17 years. The sample was stratified by age, gender, race, ethnicity, geographic region, and parent education according to US Census reports and closely matches the US population characteristics on the variables used. In addition to administration of the CAS, most of the standardization sample was also administered several achievement tests from the Woodcock–Johnson Tests of Achievement. (Woodcock & Johnson, 1989). This provided for both validity evidence and analysis of the relationships

between PASS and achievement. No further data were available on the CAS when this chapter went to press.

4.08.3 INSTRUMENT INTEGRATION

It is to the advantage of clinical and school psychologists that there are so many instruments available to assess a child's, adolescent's, or adult's intellectual functioning. Often when one instrument is administered, such as the WISC-III or WAIS-R, and then analyzed, the examiner will find that questions and hypotheses are raised regarding the individual's functioning in specific areas. Creativity and a bit of detective work are required to uncover exactly where a person's true deficits and strengths lie in their cognitive abilities. Part of the detective work in this process involves the integration of information from various instruments to support or clarify hypotheses raised as initial results are examined. Thus, examiners ultimately have to be able to integrate data from multiple instruments. As suggested by Kaufman (1994) "crucial educational decisions are sometimes made on the basis of a psychological evaluation, and these decisions should be supported by ample evidence" (p. 326) so that initial hypotheses are verified. This section describes and discusses several cognitive tests in terms of their value when integrated with results from the Wechsler Scales.

4.08.3.1 K-ABC Integration with Wechsler Scales

The K-ABC measures some of the same abilities as the WISC-III and WPPSI-R, but also measures ability in ways that are different from the Wechsler scales, thereby contributing unique information about a child's cognitive functioning. The K-ABC Simultaneous Processing Scale is believed by some researchers to involve the same cognitive requirements as Wechsler's Performance Scale (Das, Naglieri, & Kirby, 1994) and by others to be a measure of Visual Processing (Gv) (Horn, 1991). However, two Simultaneous subtests (Matrix Analogies and Photo Series), involve more reasoning (and load on two of Woodcock's (1990) factors: Fluid Reasoning (Gf) and Visual Processing (Gv)) than Wechsler's Performance subtests: The Sequential Processing Scale of the K-ABC is an excellent addition to the Wechsler because it measures sequential processing more efficiently than any Wechsler subtest. That is, because the only Wechsler test that can be viewed as measuring sequential processing is Digit Span Forward (Das et al., 1994) but the Digit Span

subtest score includes Backwards span which involves more than sequential processing (Schofield & Ashman, 1986), there is no efficient measure of sequential processing on the Wechsler. In addition, the K-ABC Achievement Scales are highly related to the Wechsler Verbal IQ and crystallized abilities (Kaufman & Kaufman, 1983; Naglieri & Jensen, 1987).

Given the above characteristics of the K-ABC, examiners may note that the entire Simultaneous Processing Scale serves as a good measure for children with motor and/or speed problems who earn low Wechsler Performance IQs, because it minimizes both of these variables. From the Horn view, the K-ABC offers good supplemental subtests to measure Gc including Faces and Places and Riddles, in the Achievement Scale. These tasks measure range of general knowledge by identifying visual stimuli (Faces and Places), and require the child to use verbal reasoning to demonstrate word knowledge (Riddles). This is unlike many tests, such as WISC-III Vocabulary, and similar Binet IV and DAS tasks, which measure word knowledge by requiring a child to retrieve word definitions from long-term storage. Like the WJ-R crystallized subtests, Riddles requires a one-word response; it is, therefore, a good Wechsler supplement to help discern whether a low V-IQ is due more to conceptual problems or expressive difficulties. The K-ABC offers some alternative modalities of receiving input and expression of response, to supplement Wechsler subtests that mainly use the auditory-vocal and visual-motor channels of communication. The K-ABC offers three subtests which call for use of the visual and vocal modalities (Magic Window, Faces and Places, and Gestalt Closure), and has one subtest that uses the auditory-motor channel (Word Order). In addition, the K-ABC taps the visual and semantic-motor channels with Reading Understanding, which requires a child to read a stimulus and do what it says (i.e., "Stand up").

4.08.3.2 Integration of KAIT with Wechsler Scales

The KAIT was developed from the Horn-Cattell theory and yields both a Crystallized IQ and Fluid IQ. The three subtests comprising the KAIT Fluid Scale are very good supplements to the Wechsler scales. As noted previously, there is controversy regarding how well the Wechsler scales measure fluid abilities; thus, it is wise to administer supplemental tests to tap an individual's fluid reasoning ability and learning ability. Assessment of planning ability, formal operational thought, and learning ability may

be obtained through KAIT Fluid Subtests: Mystery Codes and Logical Steps. Problem-solving through verbal reasoning and verbal comprehension is required in Logical Steps, and Rebus Learning demands vocal responding; therefore, the KAIT Fluid Scale measures an ability that is quite different from Wechsler's P-IQ. If questions about an individual's planning speed arise from the primary battery administered, examiners may administer Mystery Codes to further assess planning speed.

To supplement Wechsler's Verbal Scale, the KAIT Crystallized subtests may be used. For assessing an individual's base of general factual knowledge, Famous Faces may be administered to supplement WISC-III or WAIS-R Information. Famous Faces uses pictorial stimuli integrated with verbal clues about famous people; whereas Information is a purely auditory-vocal task. Formal operational thought within the Crystallized domain can be assessed through Double Meanings. Double Meanings challenges examinees to unify apparently disparate semantic stimuli. KAIT subtests, Double Meanings and Definitions, can also provide follow-up to questionable performance on tasks of word knowledge and verbal concept formation, such as Wechsler's Vocabulary or Similarities. Auditory Comprehension can be used for questions regarding an individual's memory and comprehension ability. This subtest mimics a real-life situation, in requiring an individual to listen to a mock news broadcast and answer questions about it. The two delayed recall (TSR) KAIT subtests are also very good WISC-III supplements when hypotheses are raised regarding an individual's memory.

4.08.3.3 Integration of Binet IV with Wechsler Scales

As noted in the earlier discussion of the Binet IV, there is controversial and weak factor-analytic support for the four Binet IV area scores (Verbal, Abstract-Visual, Quantitative, and Short-term Memory). The relationship between Wechsler's Verbal and Performance IQs and the Binet IV Area scores is not clear-cut. In a correlational analysis with the WISC-R and Binet IV (Thorndike et al., 1986), the WISC-R and Binet IV Verbal scales were found to relate substantially to each other. However, Kaufman (1994b) notes that the Absurdities subtest probably lowered the relationship with the Verbal IQ and increased the correlation with the Performance IQ because it uses visual stimuli. In Woodcock's (1990) factor analysis, the Binet IV Quantitative subtests loaded on a

separate Quantitative Factor (Gq), which also included Wechsler Arithmetic and WJ-R math achievement subtests. The Binet IV Abstract-Visual subtests, except Matrices, were on the Gv factor, along with most Wechsler Perceptual Organization subtests. Matrices, however, had a substantial loading on the Gf factor; it is, therefore, an excellent addition to Wechsler's Performance Scale.

The Binet IV can be integrated with Wechsler results, and can be especially helpful in assessing young children and mentally retarded individuals because of the extension of its norms down to age two. Response time is relatively unimportant on the Binet IV; therefore, it provides several subtests to further evaluate hypotheses regarding a low score on the WISC-III P-IQ or PO Index. If poor fluid intelligence is suspected, Pattern Analysis, Paper Folding and Cutting, Matrices, and Number Series can be administered. To further assess the comprehension knowledge ability measured by the WISC-III, without requiring verbal comprehension, Absurdities is especially good to administer because the stimulus is visual and minimally verbal. For assessing whether a child's fluid reasoning ability generalizes to number manipulation activities, the two Binet IV Quantitative subtests are useful. The tasks not included on the Wechsler scales such as Matrices, Equation Building, Number Series, and Verbal Relations, can be used to further explore the reasoning abilities of an individual.

4.08.3.4 Integration of WJ-R with Wechsler Scales

Wechsler's Verbal IQ primarily can be viewed as a measure of crystallized intelligence and short-term memory; Performance IQ as a blend of fluid reasoning, visual processing, and processing speed (Kaufman, 1994b). However, some researchers view Wechsler's Perceptual Organization to be a measure of visual processing (McGrew & Flanagan, 1996; Woodcock, 1990). Long-term retrieval is not specifically measured by the Wechsler scales nor is auditory processing. And, if Woodcock and others are correct, then fluid reasoning also is not measured very well by the Wechsler scales. Therefore, WJ-R Cognitive subtests provide excellent tasks for extending assessment from Wechsler subtests using the WJ-R tests which were developed to reflect Horn's pure factors. The WJ-R provides subtests which are controlled learning tasks, allowing the assessment of a person's learning ability. Conventional intelligence tests, including Wechsler's, do not typically measure this ability. The controlled

learning subtests include the following: Memory for Names and Visual–Auditory Learning (both Long-term Retrieval tasks), and Analysis–Synthesis and Concept Formation (both Fluid Reasoning tasks). Whereas the Wechsler Performance subtests emphasize visual-motor coordination and speed of response, Analysis–Synthesis and Concept Formation involve no motor coordination at all, and speed of response is not a major variable in determining a person's performance level.

The WJ-R provides several subtests from which to choose, so when questions arise regarding Wechsler's Perceptual Organization construct (including visual processing and fluid abilities, as noted above), the WJ-R is an excellent tool to investigate these hypotheses. The different aspects of the information-processing model are measured by four WJ-R factors including Gv (input), Gf (integration), Glr (storage), and Gs (output). A high or low score on Wechsler's Performance scale should be explored further to determine what aspects of an individual's information-processing may have affected this asset or deficit. If an individual is suspected of having a deficit or strength in their nonverbal visual-spatial ability, requiring further testing for clarification, then WJ-R Spatial Relations is a useful tool to make that determination. The other Fluid Reasoning subtests, previously mentioned, have a heavy verbal component and do not assess visual-spatial skills, although they do use figural material. (One precaution to note is that cognitive tests in the WJ-R battery are heavily entrenched in the tradition of measuring intelligence through predominantly verbal means (Kaufman, 1990)).

For assessing strengths or weaknesses within the auditory-vocal channel, the following WJ-R factors may be used: Ga (input), Gc (integration), and Gsm (storage). These factors can be helpful in clarifying questions raised in the Verbal scale of the Wechsler test. On the WJ-R, two Gc tasks require one-word responses, which make them good when you do not know if a low Verbal score reflects poor concepts or poor expression. Auditory-perceptual tasks on the Ga scale assess whether a child can perceive words in isolation (through filling in the gaps or by blending sounds). However, for assessing a processing deficit of longer auditory input, additional subtests may be needed (such as the Cognitive Assessment System subtests Verbal-Spatial Relations, Sentence Repetition, and Sentence Questions).

Wechsler scales do not assess long-term memory over the period of a few minutes, although they do measure short-term memory with Digit Span and remote memory with

Picture Completion and Information. The Long-term Retrieval subtests of the WJ-R provide a good assessment of the long-term memory function; therefore, these tasks complement the Wechsler subtests for supplementary analysis. In addition, the WJ-R tests provide a measure Auditory Processing and Visual Processing, which have strong perceptual components. These perceptual processes are not typically evaluated in most tests of intelligence, but need to be assessed in cases with possible neuropsychological difficulties.

4.08.3.5 DTLA-3 Integration with Wechsler Scales

The DYLA-3 has several theoretical underpinnings, including models such as fluid and crystallized intelligence, simultaneous and successive processes, and verbal and performance abilities. DTLA-3 subtests may be used to augment the Wechsler scales in several instances. To further assess perceptual organization ability, fluid ability, and simultaneous processing, Design Reproduction or Symbolic Relations may be administered. For hypotheses regarding similar fluid abilities, but tapping sequential processing, examiners may administer Design Sequences. DTLA-3 Design Reproduction is also a good supplement Wechsler Performance subtests if there is a question regarding a person's ability being hampered by response speed tests. This test does require visual-motor coordination but places minimal demands on speeded performance. Like the K-ABC, the DTLA-3 offers some alternative modalities of receiving input and expression of response to supplement Wechsler subtests that mainly use the auditory-vocal and visual-motor channels of communication. The DTLA-3 offers two subtests which call for use of the visual and vocal modalities (Story Construction and Picture Fragments), and has one subtest that uses the auditory-motor channel (Reversed Letters).

4.08.3.6 Integration of DAS with Wechsler Scales

The six Core subtests of the DAS create three separate scales for children, namely: Verbal, Spatial, and Nonverbal Reasoning. The WISC-III Verbal Comprehension subtests (specifically Vocabulary and Similarities) have been noted to be quite similar to the DAS Verbal Scale (Kaufman, 1994). The DAS Verbal, Spatial, and Nonverbal Reasoning scales have been shown to correspond to the Woodcock–Johnson Revised factors of Gc, Gv and Gf, respectively (McGhee, 1993).

The two subtests comprising the DAS Nonverbal Reasoning Scale provide an excellent addition to the WISC-III because they are quite different from WISC-III subtests. The Nonverbal Reasoning subtests (Matrices, and Sequential and Quantitative Reasoning) measure nonverbal reasoning without time limits but they do require visual-motor coordination, and minimize visualization. Thus, they can provide good measures of an individual's pure fluid ability. DAS subtests requiring visual-motor coordination, but placing minimal demands on speeded performance include Recall of Designs and Pattern Construction (when the latter test is administered via special procedures). Thus, these subtests can be useful in following up hypotheses generated from Wechsler's Performance subtests, which reward quick performance.

4.08.3.7 Integration of CAS with Wechsler Scales

Like the other tests included in this chapter, the CAS has some overlap with the WISC-III, but because its conceptualization is based on the PASS theory unique information about a child's cognitive functioning can be obtained. The CAS Planning and Attention Scales require processes that can not be effectively assessed by the Wechsler Scales (Das et al., 1994). In order to measure planning adequately, tests that evaluate the child's ability to decide how to solve problems, and determine their effectiveness are required. This means that the child must be given the opportunity to compete tasks in planful ways, unencumbered by rules imposed by the test. Additionally, items that are influenced by the child's plan rather than other factors (e.g., spatial or verbal skills) are needed. Tests of this type are not found on the WISC-III. Attention tests should demand the focus of cognitive activity and selective attending to particular information while avoiding distraction. Carefully constructed measures of attentional processes are not included on the Wechsler, yet this, as well as planning processes, are especially important when evaluating children, especially those with attention deficits and learning disabilities, for example. The measurement of Planning and Attention offer important cognitive functions that extend beyond the WISC-III and therefore offer additional information for diagnosis as well as intervention (Das et al., 1994).

The CAS, like the K-ABC, provides a measure of simultaneous processing that is similar to the demands of Wechsler's Performance Scale (Das et al., 1994) but there are

important distinctions. The CAS offers a verbal test of simultaneous processing (Verbal-Spatial Relations), one that involves memory (Figure Memory), and one with complex demands (Nonverbal Matrices). The addition of the Verbal-Spatial Relations subtest is important because it integrates both nonverbal and verbal stimuli for the comprehension of logical grammatical sentences. Similarly, the Successive processing Scale of the CAS provides tests that demand immediate recall of information (Word Series), and also measures that demand comprehension of syntax (Sentence Repetition and Sentence Questions) and where the involvement of immediate memory is markedly reduced (Successive Speech Rate).

The CAS offers a view of ability that reduces the influence of language and achievement, and therefore, provides additional information from the WISC-III or the WPPSI-R to evaluate the performance of children who are bilingual or whose educational history is problematic. Because the CAS does not have achievement or language based tests like the Wechsler (e.g., Arithmetic or Vocabulary) the reduction in the involvement of acquired knowledge provides an opportunity to evaluate children whose poor school history or language difference may have lowered their Wechsler scores. In such a situation the CAS scores can assist the psychologist in determining the extent to which low Wechsler scores may reflect language/achievement issues rather than low intellectual ability.

4.08.4 FUTURE DIRECTIONS

Intellectual assessment has changed a great deal in the twentieth century. It has moved from assessments based on language and speech patterns to sensory discrimination, with most early assessments being for the mentally deficient. Gradually, the assessment instruments developed into the precursors of the standardized instruments used in the 1990s, which measure more complex cognitive tasks for all levels of cognitive ability. The most commonly used tests, the Wechsler scales, were not developed out of theory, but were guided rather by clinical experience. In the progression of test development, the relative alternates to the Wechsler have been more theory-based. The direction in test development in the 1990s seems to be continuing to lead to theory being at the base of intelligence tests, instead of being just purely clinically driven. Psychometric theories and neurological theories are growing as the basis for new instruments. However, despite this proliferation of new theory-driven tests, there is a definite conservatism that holds on to the past,

reluctant to let Wechsler tests be truly rivaled. Part of this hold on the past is research-based and part is clinically-based. Because the Wechsler tests have been reigning supreme for so long, there has been a mountain of research studies using the Wechsler scales. Thus, clinicians have a good empirical basis to form their understanding of what a specific Wechsler profile may be indicating. Clinically, psychologists are also quite comfortable and familiar with the Wechsler scales. A good clinician who has done many assessments may be familiar enough with every nook and cranny of the WAIS-R to barely need the manual to administer it. Thus, the field so far has changed relatively slowly. Computer based technology is likely going to ultimately shape the field of assessment by 2020. Computer scoring programs and computer assisted reports are already in use, and the future is likely to include a much greater progression of technologically advanced instruments for assessing intelligence.

4.08.5 SUMMARY

This chapter first introduces the assessment of intellectual ability through a brief overview of its development throughout history. Some important faces in the history of assessment include those such as Jean Esquirol, Edouard Seguin, Sir Francis Galton, James McKeen Cattell, Alfred Binet, Lewis Termin, and David Wechsler. The early pioneers in assessment focused mainly on tests for the mentally deficient, but more recently in history tests were developed for assessing all levels of intellectual functioning. The progression of test development to the standardized instruments we know today is reviewed in the beginning of this chapter.

The debate over intelligence testing is also discussed in this chapter. Three groups of critics are presented. One controversy is raised by those opposed to subtest interpretation advocated by Wechsler; another is those that insist that all the Wechsler scales measure is *g*, rendering the different scales meaningless; and a final controversy is raised by a group who complains that the instruments do not enhance remedial interventions. Kaufman's (1979, 1994b) intelligent testing approach is presented in response to the criticism. The aim of the basic principles of intelligent testing presented in this chapter is to encourage clinicians to approach the task of profile interpretation using their knowledge of research, theory, and clinical skills, rather than being overly dependent on specific scores.

With this important philosophy of intelligent testing in mind, 10 measures of child, adoles-

cent, and adult intelligence are discussed. The Wechsler scales are generally used by examiners as the primary instrument in an assessment battery. However, there are multiple other excellent instruments available for assessing cognitive ability. The following instruments are discussed: WPPSI-R, WISC-III, WAIS-R, K-ABC, KAIT, Binet-IV, WJ-R Tests of Cognitive Ability, DTLA-3, DAS, and CAS. For each, the theory or theories underlying the instrument is presented, followed by the standardization and properties of the scale, including research using the scale, and each is completed with an overview of the instrument.

One of the main principles of the intelligent testing philosophy discussed is that hypotheses generated from the profile of the main assessment instrument should be supported with data from multiple sources. Accordingly, this chapter discusses how to supplement the Wechsler scales with additional cognitive tasks to support or clarify hypotheses raised from the initial cognitive battery. It is important for examiners to be knowledgeable and insightful in choosing supplemental measures to uncover exactly what an individual's cognitive strengths and weaknesses are, in order to form the basis for good recommendations. Specific examples are given for how examiners may supplement the Wechsler scales with each cognitive instrument presented. For example, hypotheses raised in the Wechsler profile regarding fluid reasoning may be further assessed by WJ-R subtests or KAIT subtests. If questions arise regarding the effect of response speed, the Binet IV or K-ABC Simultaneous subtests may be useful supplements. The DAS provides good supplemental information about nonverbal reasoning ability. Alternative modalities of receiving input and expression of response may be assessed by the DTLA-3 or K-ABC subtests. The K-ABC Achievement subtests may also provide additional information about an individual's verbal ability or crystallized abilities. Unique information about a child's planning ability may be found by supplementing a battery with the CAS. The multiple pieces of evidence provided by the supplemental tests, can be carefully integrated to confirm or deny hypotheses raised in an individual's cognitive profile. An illustrative case report is presented at the conclusion of this chapter, which provides an example of how many different sources of evidence are combined to provide a clear description of a 13-year-old female's cognitive and academic functioning.

The future direction of intellectual assessment appears to be in theory-driven instruments. The psychological community has held tightly onto the clinically-based Wechsler scales because of

their large research base, clinical familiarity, and just tradition. Thus, psychometric and neurological theories are becoming an important base for the newer instruments of assessment, although their use by the majority of clinicians has only slowly been occurring. Technological advances are also eminently going to be further impacting the field of assessment. The use of computer science and neurological measurement will likely begin to change the face of cognitive assessment within the next few decades.

4.08.6 ILLUSTRATIVE CASE REPORT

The following case report is of Laura S., a 13-year-old seventh grader, experiencing difficulty on standardized tests in the areas of vocabulary, reading comprehension, and writing mechanics. This report illustrates the methods and procedures for test integration and interpretation described earlier in this chapter. Table 12 provides the specific scores earned by Laura on each instrument administered.

4.08.6.1 Referral and Background Information

Laura was brought in for an evaluation by her parents, Linda and Rob (Mr. & Mrs. S) who were referred for an assessment by her current school, due to concern about the inconsistency between Laura's high grades at school and her lower standardized test scores in some areas. From Laura's scores on the Educational Record Bureau's (ERB) Comprehensive Testing Program (CTPII), the main area of concern to Laura's parents was in her Verbal Ability, specifically, Vocabulary, Reading Comprehension, and Writing Mechanics. Mr. and Mrs. S. would like to gain a better understanding of Laura's difficulty with her mathematics courses, reading comprehension, and memory retention. Laura's parent's reported that she does not always seem to understand written instructions and asks a lot of questions. They would specifically like recommendations to help Laura improve her learning ability, enhance her memory, and develop a greater aptitude for understanding directions.

Laura is a 13.5-year-old adolescent girl who has lived at home with her mother and step-father since age three. At age 2.5., her mother and biological father separated, and, other than for one month of her life, she has lived with her mother. Laura also has a 20-year-old step-sister, Kelly, and a 23-year-old step-brother, Rick, who do not live with her at home. Mr. and Mrs. S. both work full-time as professionals outside the home. Laura's biological father lives out of

state and Laura visits him once or twice a year. Laura has also had other adults serve as caretakers for her. From ages 1–3, her mother indicated that two other people helped care for her, and from ages 4–13, five housekeepers also provided care for her.

Mrs. S. reported that she had a normal, full-term pregnancy with Laura. There were no problems during the birth, and Laura was born weighing six pounds, nine ounces after a short labor of only 15 minutes. Laura's medical history is relatively unremarkable, with her parents noting only that she experienced a bilateral hernia at nine months and had several earaches and colds as a younger child. Mr. and Mrs. S. stated that their daughter has never been hospitalized, has had no major injuries, or diseases. All of Laura's developmental milestones were "on time" or "early" according to her mother.

A medical question of Laura's parents had been her hearing ability, due to her noted difficulty sometimes distinguishing particular sounds in school and occasional single-word substitutions. After discussing this at the current evaluation's intake interview, Mr. and Mrs. S. took Laura to a physician at a local university's Medicine Ear Institute to rule out any potential hearing loss. According to a letter sent by the physician, he performed a physical exam and audiogram for Laura. He stated that her audiogram "revealed normal thresholds bilaterally with excellent speech discrimination," and "she appears to have normal hearing."

Laura's educational history began when she entered preschool at age three. According to her parents, she had no difficulties with beginning preschool or transitioning to her next school. All of Laura's report cards from first through seventh grades indicate consistent, excellent achievement. Most of her teachers commented on her conscientious and serious approach to learning, and express their great pleasure having her as a student.

A classroom observation of Laura was performed for this evaluation. Laura was observed in her Honors math class. Laura was very friendly, chatting with her friends and brightly greeting the teacher right before class. However, as soon as the class began, she got right to work. When the instructor asked questions of the class, Laura raised her hand to every question. She was focused on the work all throughout the class, and seemed motivated to do well in math. She asked the teacher for assistance several times during the observation. In an interview, the teacher stated that Laura frequently asks for extra help, but it is her teacher's belief that this is mainly because "she

Table 12 Laura: Psychometric summary.

WISC-III					
Scale	IQ	Percentile rank	Factor	Index	Percentile rank
Verbal scale	114±5	82	Verbal comprehension	120±5	91
Performance scale	95±5	37	Perceptual organization	94±6	34
Full scale	106±4	66	Freedom from distractibility	101±8	53
			Processing speed	111±7	77
Subtest	Scaled score	Percentile rank	Subtest	Scaled score	Percentile rank
Information	11	63	Picture Completion	8	25
Similarities	14	91	Coding	14	91
Arithmetic	8 -W	25	Picture Arrangement	8	25
Vocabulary	12	75	Block Design	11	63
Comprehension	17 -S	99	Object Assembly	9	37
Digit span	12	75	Symbol Search	10	50
KAIT					
	IQ	Percentile rank (age)			
Fluid scale	120±5	91			
Subtest	Standard score	Percentile rank (age)			
Rebus Learning	16	98			
Logical Steps	12	75			
Mystery Codes	13	84			
WJ-R					
	Standard score	Percentile rank			
Broad reading	118	89			
Basic reading skills	125	95			
Reading comprehension	113	81			
Letter word identification	125	95			
Passage comprehension	106	65			
Word attack	117	87			
Reading vocabulary	118	88			
Broad mathematics	111	77			
Basic math skills	115	85			
Calculation	117	86			
Applied problems	102	56			
Quantitative concepts	108	70			
Broad written language	105	64			
Basic writing skills	97	41			
Written expression	130	98			
Dictation	95	37			
Writing samples	123	94			
Proofing	99	48			
Writing fluency	134	99			
Punctuation and capitalization	97	41			
Spelling	100	50			
Usage	98	44			
Broad knowledge	99	48			
Science	97	42			
Social studies	96	40			
Humanities	110	74			

Table 12 (continued)

<i>WJ-R (intra-achievement discrepancies)</i>	<i>Actual standard score</i>	<i>Predicted standard score</i>	<i>Standard deviation difference</i>	<i>Percentile rank</i>
Broad reading	118	105	+ 1.48	93
Broad mathematics	111	106	+ 0.52	70
Broad written language	105	109	− 0.45	32
Broad knowledge	99	113	− 1.20	12

does not trust her gut.” The teacher said that Laura always begins the class with a happy, cheerful mood, but she tends to “stress out” when problems become difficult.

In discussing family history related to Laura’s difficulties, Mrs. S. indicated that she feels she also has “a poor memory and retrieving skills.” Mr. and Mrs. S. noted that Laura has several strengths, including her strong intuitive abilities, her confidence, and her persistence. Laura is very popular with her peers and is viewed as a leader by many of her teachers.

In an interview with Laura, she stated that she does not enjoy reading but likes using her creativity in writing. She said that she takes school more seriously than most of her peers, and has set very high standards for herself. She explained that she is harder on herself than her parents are, especially when it comes to academics. In her free time Laura works diligently on her homework and enjoys playing tennis and soccer, and socializing with her friends.

4.08.6.2 Appearance and Behavioral Characteristics

Laura is a mature, pretty, 13-year-old seventh grader. She wore her thick brown hair parted stylishly down the middle. For each of the evaluation sessions she dressed casually and neatly in jeans and a sweatshirt. She talked comfortably when asked questions by the examiner and spoke in soft voice with an air of politeness and good manners. She tended to respond with short phrases rather than in complete sentences when answering questions posed by the examiner. However, this did not detract from her ability to articulate her thoughts clearly.

She seemed eager to please the examiner, often asking for clarification about what was the right thing to do during a certain subtest. For example, during a task which required her to look at a picture and tell what important part was missing, she asked, “Can I tell you what’s wrong with it?” Her frequent questions to the examiner indicated not only her anxiousness to

please, but also her uncomfortableness with ambiguity in a situation. She appeared much more relaxed when the situation was structured and she knew clearly everything that was expected of her.

Laura demonstrated a strong ability to concentrate and focus. She had great stamina throughout the mentally challenging evaluation. She was persistent in always refusing breaks offered to her and worked straight through during each session, displaying extreme self-control. She was cooperative and friendly, often helping the examiner put away stimulus materials and always following directions. Encouragement from the examiner was welcomed warmly with a smile by Laura. She gradually became less cautious in her casual conversation with the examiner as the testing progressed. She began to share bits and pieces of her life and showed a well-rounded self.

Laura always tried her best. Even after having attempted problems that were difficult for her, she did not lose her motivation to keep trying. In solving problems she worked quickly, but was reflective. She would continue to check her work, even after she was done, always being careful in her responses. Laura expressed anxiety about having to solve mathematical problems mentally, saying, “I can’t do things in my head, without pencils and things.” She was more tentative in answering such questions. On more difficult nonverbal tasks, Laura tended to analyze the situation first and then proceed with the task. Again, evidence of her trying carefully to do her best and not make mistakes. On the basis of her behavior during the evaluation these results should be considered a valid representation of her ability.

4.08.6.3 Tests Administered

(i) Wechsler Intelligence Scale for Children-3rd edition (WISC-III)

(ii) Woodcock–Johnson-Revised (WJ-R): Tests of Achievement

(iii) Kaufman Adolescent and Adult Intelligence Test (KAIT): Fluid Subtests

(iv) Rotter Incomplete Sentences.

4.08.6.4 Test Results and Interpretation

Laura was administered a series of cognitive tests to assess her information processing abilities. According to the WISC-III, Laura is functioning currently within the Average to High Average range of intelligence. She obtained a Verbal IQ score of 114 ± 5 (82nd percentile), Performance IQ score of 95 ± 5 (37th percentile), and Full Scale IQ score of 106 ± 4 (66th percentile). In addition, she also obtained a Verbal Comprehension Index of 120 ± 5 , which was significantly higher than her Freedom from Distractibility Index of 101 ± 8 . However, this difference within her verbal scale was due mainly to her difficulty computing arithmetic problems mentally. Laura's Processing Speed Index of 111 ± 7 was significantly higher than her Perceptual Organization Index of 94 ± 6 , indicating that she performed better on tests of visual processing speed than on tests of nonverbal reasoning.

It is important to note that Laura's Verbal subtest scores exhibit a significant amount of scatter, suggesting that her Verbal Comprehension and Perceptual Organization Indices provide a more meaningful picture of her abilities than the overall Full Scale IQ, Verbal IQ, or Performance IQ. The variance in her Verbal Scale indicates that some of her abilities are more well developed than others. The difference between her verbal and nonverbal abilities, as reflected by her Factor Indices, is unusually large and statistically significant. Because there is a 26-point difference in favor of her Verbal Comprehension Index over her Perceptual Organization Index (occurring in less than 5% of normal children), her Full Scale IQ should not be used as an indication of her overall ability. A fuller and clearer picture of Laura's abilities can be found by looking at her performance in individual areas rather than considering the statistical average of these various abilities.

In the verbal area, Laura demonstrated a significant weakness in Arithmetic, earning a score in the 25th percentile. The WISC-III arithmetic does not allow use of paper and pencil to do computation, so Laura was required to manipulate the numbers mentally to solve auditorally and visually presented problems. She expressed several times during this test that she needed to see the numbers and have them down concretely in front of her to figure the problems out. Her weakness on WISC-III Arithmetic was in contrast to her higher scores in the mathematics area on Woodcock-Johnson-Revised Tests of Achievement (WJ-R). All of the mathematics on the WJ-R allowed the use of paper and pencil in solving

the problems, which Laura was more comfortable with. Her performance on problems of calculation, such as addition, subtraction, multiplication, and division with multiple digit numbers, decimals, and fractions, was significantly better than her performance on applied problems requiring her to use mathematics to solve problems involving scenarios with money, distance, and weight. On the Woodcock-Johnson-Revised Mathematics Subtests, she scored at the 86th percentile on calculation and 56th percentile on problems that were applied. Thus, her weakness in Arithmetic on the WISC-III is not due to poor computation ability, but rather due to the fact that she needs the concrete visual stimulus of written numbers in order to utilize the mathematical knowledge that she does have.

Laura's variability in her performance on the Verbal Scale of the WISC-III was accentuated by her extremely high score (99th percentile) on a test of social judgment, verbal reasoning, and practical knowledge. Her strong verbal ability was apparent during this subtest, as well as on a subtest requiring her to use abstract reasoning to describe how two things are similar. This strength was also paralleled by her excellent performance on WJ-R tests of written expression, on which she earned an overall score at the 98th percentile. She can come up easily with vocabulary to express her ideas and is able to formulate alternative ways to get her point across if it is unclear. However, the mechanical details of written expression are more difficult for Laura. For example, she earned a lower score on a test of dictation (37th percentile), which examined her spelling, punctuation, and word usage. Laura's performance in spelling, punctuation, and word usage was at a lower level than her overall Written Expression ability (41st, 50th, and 44th percentiles, in contrast to 98th percentile). Thus, she is able to get her verbal ideas across, but is lacking skill in the grammatical rules and details of written expression.

Laura's verbal reasoning ability also appeared stronger than her knowledge of general factual knowledge. She scored at the 63rd percentile on a WISC-III task requiring her to answer questions about information acquired from formal schooling. However, her performance on WJ-R tests of Broad Knowledge was lower than expected, given her high level of academic achievement at school, as well as her performance on other WJ-R tests of achievement. On subtests covering areas such as science, social studies, and humanities, Laura scored in the Average level at the 48th percentile. A person with Laura's total achievement performance would have been expected to

earn a slightly higher score on these tests of Broad Knowledge. Her actual standard score was 1.2 standard deviations lower than what was predicted in this area, indicating that she is not achieving at a level consistent with what would be expected given her level of achievement. This is also reflected in her lower ERB group standardized test scores that are discrepant from her higher school grades. The ERB scores reflect strictly facts derived from multiple choice exams. However, her interaction with teachers, participation and performance at school allow the teachers to understand Laura as a more complete person, which positively influences her grades.

Laura's performance on WJ-R Broad Reading (89th percentile) is also commensurate with her strong verbal reasoning skills evidenced. However, her score on a task measuring comprehension of a written passage was somewhat lower than expected (65th percentile) given her strong verbal abilities. Nonetheless, it is in the average range and not low enough to be of concern. Overall, Laura's strong performance on Broad Reading ability was 1.48 standard deviations above what would be predicted for an individual with her total achievement performance, so she is demonstrating strong ability with her reading skills. Only 7% of students Laura's age, who had the same expected standard score as she, scored as high or higher than Laura on Broad Reading Subtests of the WJ-R.

Laura's significantly higher score on the Verbal Comprehension Index than the Perceptual Organization Index of the WISC-III indicates that her verbal abilities are more well developed than her nonverbal and perceptual abilities. However, even on the nonverbal performance tests, all of her scores were at or above the Average level (25th to 91st percentiles). For example, Laura performed well on a task requiring her to use short-term memory in copying a series of different symbols from a visually presented key. She scored at the 91st percentile on this task and scored at the 63rd percentile on another nonverbal task requiring her to reproduce a model using blocks.

Her cognitive skills were further assessed by the Fluid Scale of the Kaufman Adolescent and Adult Intelligence Test (KAIT). The KAIT Fluid subtests measure one's ability to solve novel problems using reasoning, memory, paired-associate learning, verbal comprehension, and perceptual organization. Laura scored well above average (standard score 120 ± 5) on these tasks involving her ability to solve novel problems. She scored at the 98th percentile on a task that essentially required her to learn a new language through paired-associate learning. On

this task she used her good verbal concentration, expression ability and memory to succeed. She also performed quite well on tasks that required her to use abstract reasoning with novel stimuli and planning ability.

4.08.6.5 Summary and Diagnostic Impressions

Laura is a mature, pretty 13-year-old girl who was brought in for a psychoeducational evaluation by her parents due to their concern with inconsistency between her standardized tests scores and high grades at La Jolla Country Day School. Her parents' main areas of concern involve her abilities in vocabulary, reading comprehension, writing mechanics, and mathematics. Mr. and Mrs. S. wanted a better understanding of these discrepancies in her performance, as well as her difficulty with memory retention and understanding written directions.

Laura demonstrated strong motivation to please the examiner, persistence, and stamina during her evaluation. She was reflective in her problem solving and careful in her responses, trying hard to perform to the best of her ability. She earned scores in Average to High Average range of intelligence on the WISC-III. Her factor indices and specific strengths and weaknesses gave the most meaningful picture of Laura's abilities, due to the scatter in her verbal subscales. She performed significantly better on tests of verbal reasoning and word knowledge than on tests of nonverbal ability and visual-perceptual skills. She demonstrated a weakness on a task requiring her to mentally solve auditorally and visually presented arithmetic problems. However, this was due to her difficulty performing mathematical calculations in her head without the concrete visual help of pencil and paper, as evidenced by her higher scores on WJ-R tests of Mathematics that allow use of pencil and paper for problem-solving.

Her strengths were in the area of verbal reasoning. This was congruent with her strong ability to express herself in a written format on the WJ-R and also with her overall reading ability demonstrated on the WJ-R. Although her ability to express herself verbally and in writing was good, her skills in the details of writing, such as spelling, punctuation, and usage were not as strong. In addition, her performance on a test of passage comprehension was not as high as expected given her overall verbal abilities. In the area of general knowledge, including science, social studies, and humanities, Laura did not perform as well as one would predict given her other achievement scores.

On an additional cognitive test measuring Laura's ability to solve novel problems, she demonstrated well above average abilities. On the KAIT Fluid subtests, Laura's performance indicated that she has strong ability to learn new material through paired-associate learning. She also evidenced above average ability to use abstract reasoning and planning ability with novel stimuli.

Laura has learned very well to compensate for her areas of weakness, which is evident in her high grades at school. Her ability to ask questions when uncertain, and create a structured environment for herself so she is most comfortable are examples of such compensatory strategies. Additionally, this strength is reflected in her higher grades at school compared to her ERB scores. Her ability to express herself well in writing and vocally, such as on essay tests or class discussions, helps her grades at school but is not able to aid her in the cut-and-dry responses needed for the ERB tests. Her strong ability to solve novel problems will be quite useful to Laura in continuing to creatively create other strategies when new difficulties appear in her life.

4.08.6.6 Recommendations

The following recommendations have been made to assist Laura and her parents in enhancing her learning ability.

(i) Laura is a highly self-motivated student who has created very high expectations for herself, and has been working very hard to meet those high standards. At times this causes anxiety for her, and when this anxiety reaches an unmanageable level it may cause difficulty and decreased performance. It will therefore be useful for Laura and her parents to discuss that it is okay not to be perfect in every area of her life, every minute of the day, and to allow Laura to experience those instances of nonperfection. Tolerance and appreciation for her own continuum of strengths to weaknesses needs to be developed to keep Laura feeling good about herself.

(ii) As Laura works so hard at not making any mistakes, she is sometimes hesitant to proceed without asking many questions to prevent making any sort of error. She demonstrated strong abilities in many areas and therefore should be encouraged to go with her gut feeling, and try to attempt problems that normally she may immediately ask for help on. This will further foster her sense of independence and confidence in her own abilities.

(iii) Laura expressed being uninterested in some subjects that have not stressed problem-

solving or reasoning as learning methods, such as Social Studies. If her school classes present material in a factual, "memorize this" style, then Laura needs to study in more creative way to enhance her learning and reduce boredom. For example, she can make up a story about a character that may have lived in a period of history that she is learning about. As another example, Laura can draw and illustrate a timeline of important historical events to remember.

(iv) To incorporate a different method of studying subjects that are especially tedious to Laura, she could study with another diligent student. Similarly, she may want to get together with a group of peers and create a quiz show to reinforce and study facts that may seem boring to her when she is studying alone.

(v) To encourage her to become comfortable with and reinforce information that may be presented in a test format such as the ERB tests, Laura may create a competition for herself. For example, on a weekly basis she could give herself a test on a content area (such as those in books created for SAT preparation), and then she could chart her own progress from week to week.

(vi) It is important to reassure Laura that it is fine to use the compensatory strategies that work for her, such as writing down arithmetic problems. Her ability to figure out such strategies is a strength that was evident in her ability to solve novel problems and can be used to help her in areas that are more difficult for her.

(vii) To help Laura increase her overall academic abilities, Laura will benefit from broadening her base of what she considers "studying." Studying includes not only completing assignments given at school and preparing for exams, but also includes an awareness of one's environment outside of the school context. In broadening her conceptualization of studying, for example, Laura may incorporate more pleasure reading of nonschool books or magazines into her weekly routine, which will benefit her grammar and vocabulary. She may watch a movie or television program and relate it to some topic she is studying in school. This is important as people who improve their general learning, tend to do better overall on standardized types of testing.

(viii) Laura demonstrated strong written expression ability, but weaker ability to incorporate correctly details of grammar such as spelling and punctuation. These details may be less interesting to Laura, thus, again she may want to use her problem-solving ability to create more interesting ways to learn such details. There are many excellent computer programs that help teach grammatical details, in an interesting manner. As Laura tends to strive

for excellence, she may set goals for herself according to a computer program she is working with to help her.

(ix) Laura is very conscientious and focused in her school work. She may benefit and find it rewarding to become a peer tutor for a lower grade child (such as a fourth- or fifth-grade student). Tutoring a child who is having trouble with grammar will reinforce rules in Laura's mind as she gains self respect by being appointed to teach someone else.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Kristee A. Beres, Randy Kamphaus, Nadeen L. Kaufman, Jack A. Naglieri, and Mitch Perlman for their contributions to this chapter.

4.08.7 REFERENCES

- American Psychological Association (1990). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.
- Bain, S. K. (1991). Test reviews: Differential ability scales. *Journal of Psychoeducational Assessment*, 9, 372–378.
- Binet, A., & Henri, V. (1895). La psychologie individuelle. *L'Année Psychologique*, 2, 411–465.
- Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191–244.
- Binet, A., & Simon, T. (1908). Le développement de l'intelligence chez les enfants. *L'Année Psychologique*, 14, 1–94.
- Bogen, J. E. (1975). Some educational aspects of hemispheric specialization. *UCLA Educator*, 17, 24–32.
- Bracken, B. A. (1985). A critical review of the Kaufman Assessment Battery for Children (K-ABC). *School Psychology Review*, 14, 21–36.
- Braden, J. P. (1992). Test reviews: The differential ability scales and special education. *Journal of Psychoeducational Assessment*, 10, 92–98.
- Brown, D. T. (1994). Review of the Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of School Psychology*, 32, 85–99.
- Buckhalt, J. A. (1991). A critical review of the Wechsler Preschool and Primary Scale of Intelligence Revised (WPPSI-R). *Journal of Psychoeducational Assessment*, 9, 271–279.
- Canter, A. (1990). A new Binet, an old premise: A mismatch between technology and evolving practice. *Journal of Psychoeducational Assessment*, 8, 443–450.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, 1–22.
- Cohen, J. (1957). A factor-analytically based rationale for the Wechsler-Adult Intelligence Scale. *Journal of Consulting Psychology*, 6, 451–457.
- Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). *Psychological testing*. Mountain View, CA: Mayfield.
- Das, J. P. (1973). Structure of cognitive abilities: Evidence for simultaneous and successive processing. *Journal of Educational Psychology*, 65, 103–108.
- Das, J. P., Kirby, J. R., & Jarman, R. F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychological Bulletin*, 82, 87–103.
- Das, J. P., Kirby, J., & Jarman, R. F. (1979). *Simultaneous and successive cognitive processes*. New York: Academic Press.
- Das, J. P., Naglieri, J. A., & Kirby, J. (1994). *Assessment of cognitive processes*. Boston: Allyn & Bacon.
- Delugach, R. (1991). Test review: Wechsler Preschool and Primary Scale of Intelligence-Revised. *Journal of Psychoeducational Assessment*, 9, 280–290.
- Dumont, R., & Hagberg, C. (1994). Test reviews: Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of Psychoeducational Assessment*, 12, 190–196.
- Elliott, C. D. (1990). *Differential Ability Scales (DAS) administration and scoring manual*. San Antonio, TX: Psychological Corporation.
- Elliott, C. D. (1997). The Differential Ability Scales (DAS). In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Beyond traditional intellectual assessments: Contemporary and emerging theories, tests, and issues* (pp. 183–208). New York: Guilford Press.
- Evans, J. H., Carlsen, R. N., & McGrew, K. S. (1993). Classification of exceptional students with the Woodcock-Johnson Psycho-Educational Battery-Revised. In R. S. MacCallum (Ed.), *Journal of Psychoeducational Assessment monograph series. Advances in psychoeducational assessment: Woodcock-Johnson Psycho-educational Battery-Revised* (pp. 6–19). Germantown, TN: Psychoeducational Corporation.
- Flanagan, D. P., Alfonso, V. C., & Flanagan, R. (1994). A review of the Kaufman Adolescent and Adult Intelligence Test: An advancement in cognitive assessment? *School Psychology Review*, 23, 512–525.
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95, 29–51.
- Glutting, J. J., McDermott, P. A., Prifitera, A., & McGrath, E. A. (1994). Core profile types for the WISC-III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies. *School Psychology Review*, 23, 619–639.
- Goldern, C. J. (1981). The Luria-Nebraska Children's Battery: Theory and formulation. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment of the school-age child issues and procedures* (pp. 277–302). New York: Grune and Stratton.
- Hammill, D. D. (1991). *Interpretive Manual for Detroit Tests of Learning Aptitude*: (3rd ed.). Austin, TX: PRO-ED.
- Hammill, D. D., & Bryant, B. R. (1991). *Interpretive Manual for Detroit Tests of Learning Aptitude-Primary: Second Edition*. Austin, TX: PRO-ED.
- Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII* (4th ed.). Stanford, CA: Stanford University Press (Distributed by Consulting Psychologists Press).
- Horn, J. L. (1985). Remodeling old model in intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 267–300). New York: Wiley.
- Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 61–116). New York: Freeman.
- Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock-Johnson Technical manual: A reference on theory: and current research* (pp. 197–246). Allen, TX: DLM Teaching Resources.
- Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, 57, 253–270.
- Horn, J. L., & Cattell, R. B. (1967). Age difference in fluid and crystallized intelligence. *Acta Psychologica*, 26, 107–129.
- Hoy, C., Gregg, N., Jagota, M., King, M., Moreland, C., & Manglitz, E. (1993). Relationship between the Wechsler

- Adult Intelligence Scale-Revised and the Woodcock-Johnson Test of Cognitive Ability-Revised among adults with learning disabilities in university and rehabilitation settings. In R. S. MacCallum (Ed.), *Journal of Psychoeducational Assessment monograph series. Advances in psychoeducational assessment: Woodcock-Johnson Psycho-educational Battery-Revised* (pp. 54-63). Germantown, TN: Psychoeducational Corporation.
- Inglis, J., & Lawson, J. (1982). A meta-analysis of sex differences in the effects of unilateral brain damage on intelligence test results. *Canadian Journal of Psychology*, 36, 670-683.
- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence*. Boston: Allyn & Bacon.
- Kamphaus, R. W., Beres, K. A., Kaufman, A. S., & Kaufman, N. L. (1995). The Kaufman Assessment Battery for Children (K-ABC). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed., pp. 348-399). Boston: Allyn & Bacon.
- Kamphaus, R. W., & Reynolds, C. R. (1987). *Clinical and research applications of the K-ABC*. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S. (1979). *Intelligent testing with the WISC-R*. New York: Wiley.
- Kaufman, A. S. (1983). Intelligence: Old concepts—new perspectives. In G. Hynd (Ed.), *The school psychologist* (pp. 95-117). Syracuse, NY: Syracuse University Press.
- Kaufman, A. S. (1985). Review of Wechsler Adult Intelligence Scale-Revised. In J. V. Mitchell (Ed.), *The ninth mental measurement yearbook* (pp. 1699-1765). Lincoln, NE: University of Nebraska Press.
- Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence*. Boston: Allyn & Bacon.
- Kaufman, A. S. (1992). Evaluation of the WISC-III and WPPSI-R for gifted children. *Roeper Review*, 14, 154-158.
- Kaufman, A. S. (1993). King WISC the third assumes the throne. *Journal of School Psychology*, 31, 345-354.
- Kaufman, A. S. (1994a). A reply to MacMann and Barnett: Lessons from the blind men and the elephant. *School Psychology Quarterly*, 9, 199-207.
- Kaufman, A. S. (1994b). *Intelligent testing with the WISC-III*. New York: Wiley.
- Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development*, 47, 165-171.
- Kaufman, A. S., & Horn, J. L. (1996). Age changes on test of fluid and crystallized ability for women and men on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at ages 17-94 years. *Archives of Clinical Neuropsychology*, 11, 97-121.
- Kaufman, A. S., & Kamphaus, R. W. (1984). Factor analysis of the Kaufman Assessment Battery for Children (K-ABC) for ages 2¹/₂ through 12¹/₂ years. *Journal of Educational Psychology*, 76, 623-637.
- Kaufman, A. S., & Kaufman, N. L. (1983). *Interpretive manual for the Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S., & Kaufman, N. L. (1993). *Interpretive Manual for Kaufman Adolescent & Adult Intelligence Test*. Circle Pines, MN: American Guidance Service.
- Kaufman, A. S., Ishikuma, T., & Kaufman, N. L. (1994). A Horn analysis of the factors measured by the WAIS-R, Kaufman Adolescent and Adult Intelligence Test (KAIT), and two new brief cognitive measures for normal adolescents and adults. *Assessment*, 1, 353-366.
- Kaufman, A. S., Kaufman, J. C., Balgopal, R., & McLean, J. E. (1996). Comparison of three WISC-III short forms: Weighing psychometric, clinical, and practical factors. *Journal of Clinical Child Psychology*, 25, 97-105.
- Kaufman, A. S., Kaufman, J. C., Chen, T., Kaufman, N. L. (1996). Differences on six Horn abilities for 14 age groups between 15-16 and 75-94 years. *Psychological Assessment*, 8, 1-11.
- Kaufman, A. S., & Lichtenberger, E. O. (in press). *WAIS-III assessment made simple*. New York: Wiley.
- Kaufman, A. S., & McLean, J. E. (1992, November). *An investigation into the relationship between interests and intelligence*. Paper presented at Annual meeting of the Mid-South Educational Research Association, Knoxville, TN.
- Kaufman, A. S., McLean, J. E., & Lincoln, A. (1996). The relationship of the Myers-Briggs Type Indicator to IQ level and fluid-crystallized discrepancy on the Kaufman Adolescent and Adult Intelligence Test (KAIT). *Assessment*, 3, 225-239.
- Kaufman, A. S., McLean, J. E., & Reynolds, C. R. (1988). Sex, race, residence, region, and education differences on the 11 WAIS-R subtests. *Journal of Clinical Psychology*, 44, 213-248.
- Keith, T. Z. (1990). Confirmatory and hierarchical confirmatory analysis of the Differential Ability Scales. *Journal of Psychoeducational Assessment*, 8, 391-405.
- Keith, T. Z., & Dunbar, S. B. (1984). Hierarchical factor analysis of the K-ABC: Testing alternate models. *Journal of Special Education*, 18, 367-375.
- Kinsbourne, M. (Ed.) (1978). *Asymmetrical function of the brain*. Cambridge, MA: Cambridge University Press.
- Levy, J., & Trevarthen, C. (1976). Metacognition of hemispheric function in human split-brain patients. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 299-312.
- Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.
- Luria, A. R. (1973). *The working brain: An introduction to neuro-psychology*. London: Penguin.
- Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.
- MacMann, G. M., & Barnett, D. W. (1994). Structural analysis of correlated factors: Lessons from the Verbal-Performance dichotomy of the Wechsler scales. *School Psychology Quarterly*, 9, 161-197.
- Matarazzo, J. D. (1985). Review of Wechsler Adult Intelligence Scale-Revised. In J. V. Mitchell (Ed.), *The ninth mental measurement yearbook* (pp. 1703-1705). Lincoln, NE: University of Nebraska Press.
- McCallum, R. S., & Merritt, F. M. (1983). Simultaneous-successive processing among college students. *Journal of Psychoeducational Assessment*, 1, 85-93.
- McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment*, 8, 290-302.
- McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education*, 25, 504-526.
- McGhee, R. (1993). Fluid and crystallized intelligence: Confirmatory factor analysis of the Differential Ability Scales, Detroit Tests of Learning Aptitude-3, and Woodcock-Johnson Psycho-Educational Battery-Revised. In B. A. Bracken & R. S. McCallum (Eds.), *Journal of Psychoeducational Assessment monograph series, advances in psychoeducational assessment: Woodcock-Johnson Psycho-Educational Battery-Revised* (pp. 39-53). Germantown, TN: Psychoeducational Corporation.
- McGrew, K. S., & Flanagan, D. P. (1996). The Wechsler Performance scale debate: Fluid intelligence (Gf) or visual processing. *NASP Communique*, 24, 15-17.
- McShane, D., & Cook, V. (1985). Transcultural intellectual

- assessment: Performance by Hispanics on the Wechsler scales. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 385–426). New York: Wiley.
- McShane D. A., & Plas, J. M. (1984). The cognitive functioning of American Indian children: Moving from the WISC to the WISC-R. *School Psychology Review*, 13, 61–73.
- Merz, W. R. (1985). Test review of Kaufman Assessment Battery for Children. In D. J. Keyser & R. C. Sweetland (Eds.), *Test overviews* (pp. 393–405). Kansas City, MO: Test Corporation of America.
- Miller, T. L., & Reynolds, C. R. (1984). Special issue ... The K-ABC. *Journal of Special Education*, 8, 207–448.
- Naglieri, J. A. (1984). Concurrent and predictive validity of the Kaufman Assessment Battery for Children with a Navajo sample. *Journal of School Psychology*, 22, 373–380.
- Naglieri, J. A., & Das, J. P. (1988). Planning–Arousal–Simultaneous–Successive (PASS): A model for assessment. *Journal of School Psychology*, 26, 35–48.
- Naglieri, J. A., & Das, J. P. (1990). Planning, Attention, Simultaneous, and Successive (PASS) cognitive processes as a model for intelligence. *Journal of Psychoeducational Assessment*, 8, 303–337.
- Naglieri, J. A., & Das, J. P. (1996). *Das–Naglieri Cognitive Assessment System*. Chicago: Riverside.
- Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, 11, 21–43.
- Obringer, S. J. (1988, November). *A survey of perceptions by school psychologists of the Stanford–Binet IV*. Paper presented at the meeting of the Mid-South Educational Research Association, Louisville, KY.
- O'Grady, K. E. (1983). A confirmatory maximum likelihood factor analysis of the WAIS-R. *Journal of Consulting and Clinical Psychology*, 51, 826–831.
- Perlman, M. D. (1986). *Toward an integration of a cognitive-dynamic view of personality: The relationship between defense mechanisms, cognitive style, attentional focus, and neuropsychological processing*. Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.
- Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development*, 15, 1–12.
- The Psychological Corporation (1997). *Wechsler Adult Intelligence Scale-Third edition (WAIS-III)*. San Antonio, TX: Author.
- Reitan, R. M. (1955). Certain differential effects of left and right cerebral lesions in human adults. *Journal of Comparative and Physiological Psychology*, 48, 474–477.
- Reschly, D. J., & Tilly, W. D. (1993, September). The WHY of system reform. *Communique*, pp. 1, 4–6.
- Reynolds, C. R. (1987). Playing IQ roulette with the Stanford–Binet, 4th edition. *Measurement and Evaluation in Counseling and Development*, 20, 139–141.
- Reynolds, C. R., Chastain, R. L., Kaufman, A. S., & McLean, J. E. (1987). Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology*, 25, 323–342.
- Reynolds, C. R., Kamphaus, R. W., & Rosenthal, B. L. (1988). Factor analysis of the Stanford–Binet Fourth Edition for ages 2 years through 23 years. *Measurement and Evaluation in Counseling and Development*, 2, 52–63.
- Roback, A. A. (1961) *History of psychology and psychiatry*. New York: Philosophical Library.
- Sandoval, J. (1992). Test Reviews: Using the DAS with multicultural populations: Issues of test bias. *Journal of Psychoeducational Assessment*, 10, 88–91.
- Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Sattler.
- Schaw, S. R., Swerdlik, M. E., & Laurent, J. (1993). Review of the WISC-III. In B. A. Bracken & R. S. McCallum (Eds.), *Journal of Psychoeducational Assessment monograph series, advances in psychoeducational assessment: Wechsler Intelligence Scale for Children-Third edition* (pp. 151–160). Germantown, TN: Psychoeducational Corporation.
- Schmidt, K. L. (1994). Review of Detroit Tests of Learning Aptitude-Third Edition. *Journal of Psychoeducational Assessment*, 12, 87–91.
- Schofield, N. J., & Ashman, A. F. (1986). The relationship between Digit Span and cognitive processing across ability groups. *Intelligence*, 10, 59–73.
- Siehn, F. A. (1985). *Correlational study of Woodcock–Johnson deviation IQ scores and WAIS-R with adult population*. Unpublished manuscript, Arizona State University.
- Sperry, R. W. (1968). Hemisphere deconnection and unity in conscious awareness. *American Psychologist*, 23, 723–733.
- Sperry, R. W. (1974). Lateral specialization in the surgically separated hemispheres. In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program*. Cambridge, MA: MIT Press.
- Spruill, J. (1984). Wechsler Intelligence Scale-Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test overviews* (pp. 728–739). Kansas City, MO: Test Corporation of America.
- Spruill, J. (1987). Review of Stanford–Binet Intelligence Scale, Fourth edition. In D. J. Keyser & R. C. Sweetland (Eds.), *Test overviews* (pp. 544–559). Kansas City, MO: Test Corporation of America.
- Sternberg, R. J. (1993). Rocky's back again: A review of the WISC-III. In B. A. Bracken & R. S. McCallum (Eds.), *Journal of Psychoeducational Assessment monograph series, advances in psychoeducational assessment: Wechsler Intelligence Scale for Children-Third Edition* (pp. 161–164). Germantown, TN: Psychoeducational Corporation.
- Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986a). *Technical manual for the Stanford–Binet Intelligence Scale-Fourth Edition*. Chicago: Riverside.
- Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986b). *Stanford–Binet Intelligence Scale: Fourth Edition*. Chicago: Riverside.
- Turkheimer, E., & Farace, E. (1992). A reanalysis of gender differences in IQ scores following unilateral brain lesions. *Psychological Assessment*, 4, 498–501.
- Turkheimer, E., Farace, E., Yfo, R. A., & Bigler, E. D. (1993). Quantitative analysis of gender differences in the effects of lateralized lesions on verbal and performance IQ. *Intelligence*, 17, 461–474.
- VanLeirsburg, P. (1994). Review of Detroit Tests of Learning Aptitude-3. In D. J. Keyser & R. C. Sweetland (Eds.), *Test overviews* (pp. 219–225). Kansas City, MO: Test Corporation of America.
- Wada, J., Clarke, R., & Hamm, A. (1975). Cerebral hemisphere asymmetry in humans. *Archives of Neurology*, 37, 234–246.
- Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review*, 23, 640–651.
- Webster, R. E. (1994). Review of Woodcock–Johnson Psycho-educational Battery-Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test overviews* (pp. 804–815). Kansas City, MO: Test Corporation of America.
- Wechsler, D. (1939). *Measurement of adult intelligence*. Baltimore: Williams & Wilkins.
- Wechsler D. (1958). *Measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Willilams & Wilkens.
- Wechsler D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised (WAIS-R)*. San Antonio, TX: Psychological Corporation.

- Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R)*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition, (WISC-III)*. San Antonio, TX: Psychological Corporation.
- Witt, J. C., & Gresham, F. M. (1985). Review of the Wechsler Intelligence Scale for Children-Revised. In J. V. Mitchell (Ed.), *Ninth mental measurements yearbook* (pp. 1716–1719). Lincoln, NE: University of Nebraska Press.
- Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psycho-educational Assessment*, 8, 231–258.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Tests of Cognitive Ability: Standard and supplemental batteries*. Chicago: Riverside.
- Woodcock, R. W., & Mather, N. (1989). WJ-R Tests of Cognitive Ability-Standard and Supplemental Batteries: Examiner's Manual. In R. W. Woodcock & M. B. Johnson (Eds.) *Woodcock-Johnson psycho-educational battery-revised*. Allen, TX: DLM Teaching Resources.

4.09

Assessment of Memory, Learning, and Special Aptitudes

ROBYN S. HESS

University of Nebraska at Kearney, NE, USA

and

RIK CARL D'AMATO

University of Northern Colorado, Greeley, CO, USA

4.09.1 INTRODUCTION	239
4.09.1.1 <i>Assessment Approaches</i>	241
4.09.1.2 <i>Evaluation of Domain Areas</i>	243
4.09.1.3 <i>Intervention Approaches</i>	244
4.09.2 ASSESSMENT OF MEMORY	245
4.09.2.1 <i>Attention</i>	246
4.09.2.2 <i>Short-term and Long-term Memory</i>	247
4.09.2.3 <i>Memory: Implications for Intervention</i>	249
4.09.3 ASSESSMENT OF LEARNING	249
4.09.3.1 <i>Models of Learning</i>	249
4.09.3.2 <i>Learning Processes: Input and Integration</i>	252
4.09.3.3 <i>Academic Achievement: Output</i>	253
4.09.3.4 <i>Learning: Implications for Intervention</i>	254
4.09.4 ASSESSMENT OF SPECIAL APTITUDES	254
4.09.4.1 <i>Sensory Perception</i>	255
4.09.4.1.1 <i>Sensory perception: implications for intervention</i>	256
4.09.4.2 <i>Motor: Fine and Gross</i>	256
4.09.4.3 <i>Sensory-motor Integration</i>	257
4.09.4.3.1 <i>Motor: implications for intervention</i>	257
4.09.4.4 <i>Communication/Language</i>	257
4.09.4.4.1 <i>Communication/language: implications for intervention</i>	259
4.09.5 FUTURE DIRECTIONS	259
4.09.6 SUMMARY	260
4.09.7 REFERENCES	261

4.09.1 INTRODUCTION

The inner workings of the human mind and the way in which people process information has intrigued researchers for centuries. The

traditional categories of cognition, such as attention, memory, language, and learning are terms frequently used and considered all important to our daily adaptive functioning and ability to learn new information. Yet all

have defied simple explanation and manipulation (e.g., Gaddes & Edgell, 1994; Lezak, 1995). That is, as professionals we are able to identify when a child or adult is having difficulty processing information, but the exact interrelations between an individual's different capacities in areas such as attention, learning style, and sensory integration still eludes educational and clinical specialists. More puzzling still is finding effective rehabilitation strategies to address deficits in memory, learning, and other psychological processes. These areas are frequently addressed in the growing volumes of neuropsychological research, but are relevant to the practices of many traditionally trained psychologists as well (Hamsher, 1984). Not surprisingly, all of the information needed for an adequate understanding and interpretation of cognitive processes cannot be provided in a single chapter or obtained in an individual university course. Thus, the purpose of this chapter is to provide a brief description of our higher cognitive processes and introduce a variety of strategies and measures for evaluating these functions.

Problems in attention, memory, and learning are not isolated to the very young or the very old. Adult learning problems often become apparent in employment settings and after injuries resulting from strokes, accidents, or diseases. Epidemiological studies suggest that in the decade following the late 1990s there will be a dramatic increase in the number of individuals suffering from organically related disorders resulting from the abuse of alcohol and other types of toxic substances (Touyz, Byrne, & Gilandas, 1994). The recent documented increase in the number of head injuries caused by motor vehicle accidents has become an intrinsic fact in today's society. High speed transportation and the growing prevalence of violent street crimes have further increased the incidence of head injuries (Touyz et al., 1994). So too, the growing popularity of certain contact sports (e.g., hockey, boxing), noncontact sports (e.g., rock climbing, mountaineering, bicycling), and recreational activities (e.g., skateboarding, roller blading) has contributed significantly to the number of individuals suffering from traumatic brain injury (Drew & Templer, 1992; Templer & Drew, 1992). Internal processes such as eating disorders, depression, diseases, epilepsy, and tumors can result in impaired executive functioning as well (Black & Strub, 1994). These realities of today's society make it necessary for clinicians to be able to accurately evaluate a client's strengths and weaknesses in everyday functioning and find the key elements to fostering effective behavioral change through rehabilitation or educational improvement.

Psychology has made many new inroads into understanding the learning process and the subsequent development of corrective or adaptive programs for children and adults with learning disorders and traumatic brain injuries. As assessment specialists, psychologists must quickly and accurately wade through the cumulative data available about the individual in order to select the most viable of alternative hypotheses to explain the findings and offer appropriate interventions (D'Amato & Dean, 1989a; D'Amato, Rothlisberg, & Leu, in press; Gutkin & Reynolds, 1990). Although administering a test may be a routine activity, conducting a thorough, valid assessment is an extremely complex process. The clinician is required to make decisions regarding which skills to evaluate and the best instrument to use with a particular client, and to generate accurate interpretations of the results in order to create the most effective intervention plan. Adding to the immensity of this task are the wide range of client variables that can impact the assessment process, including motivation, environment, culture, age, developmental level, language, training, educational quality, personal experience, and attitude to name just a few (Golden, Sawicki, & Franzen, 1984; Hynd & Semrud-Clikeman, 1990).

The clinician must recognize that the context of the client may influence or even define the outcome of the assessment (Dana, 1993; Figueroa & Garcia, 1994). For example, many of the psychological, educational, and personality instruments available to practitioners have been criticized as culturally biased, as traditionally not including individuals from diverse ethnic backgrounds in the norming sample, and as measuring acquired knowledge rather than an individual's responsiveness to instruction or the learning process (Cole & Siegel, 1990; Dana, 1993; Figueroa & Garcia, 1994; Sattler, 1992). Because of these problems, several researchers believe that standardized assessment may have questionable validity for those clients who represent diverse cultural groups (Cole & Siegel, 1990; Dana, 1993; Figueroa & Garcia, 1994; Martinez, 1985). From the beginning, one robust assumption of standardized testing was that all individuals who take the tests would have had equal or comparable exposure to the contents of the assessment materials prior to the assessment (e.g., Colvin, 1921; Dearborn, 1921; Woodrow, 1921). In direct contrast to this supposition, current statistics indicate that the US immigrant population is not only growing rapidly but is also quickly expanding in diversity (Figueroa & Garcia, 1994). These authors conclude that tests, although given high status in US society, are actually quite fragile because

of the founding assumption regarding homogeneity and general shortcomings in technical properties. Nevertheless, standardized tests can be useful in evaluating current functioning especially when multiple sources of information and multidimensional functions are evaluated to measure individual processes (Sattler, 1992). The responsible clinician must recognize both the assets and the limitations when using standardized measures with ethnic minority clients.

The strategy the practitioner uses to accomplish an effective assessment must, of necessity, be based upon well-grounded, empirically validated theories of cognition and behavior. Only through the use of a theoretical framework are specific predictions regarding performance under a given set of ecological circumstances made possible (Dean, 1985a, 1986; Rothlisberg, 1992). Unfortunately, no single, diagnostic paradigm or theory has proven sufficient to explain fully the vagaries of behavior (D'Amato & Rothlisberg, 1992). Psychoanalytically, behaviorally, and biologically based approaches, as well as other theoretical positions, have been continually challenged not only to describe behavior, but also to provide effective interventions for the populations whom they serve (D'Amato & Dean, 1989b; Gutkin & Reynolds, 1990). Prepackaged programs dealing with psycholinguistic or visual-motor training, and sensory integration training have attempted this, but typically failed to meet the demands of this challenge. Gradually, the field has acknowledged that the effective use of assessment procedures, including educational and psychological tests, is reliant upon a theoretical foundation, which allows the incorporation of information from multiple data sources and environments in such a manner as to increase the amount of effective and appropriate interventions generated.

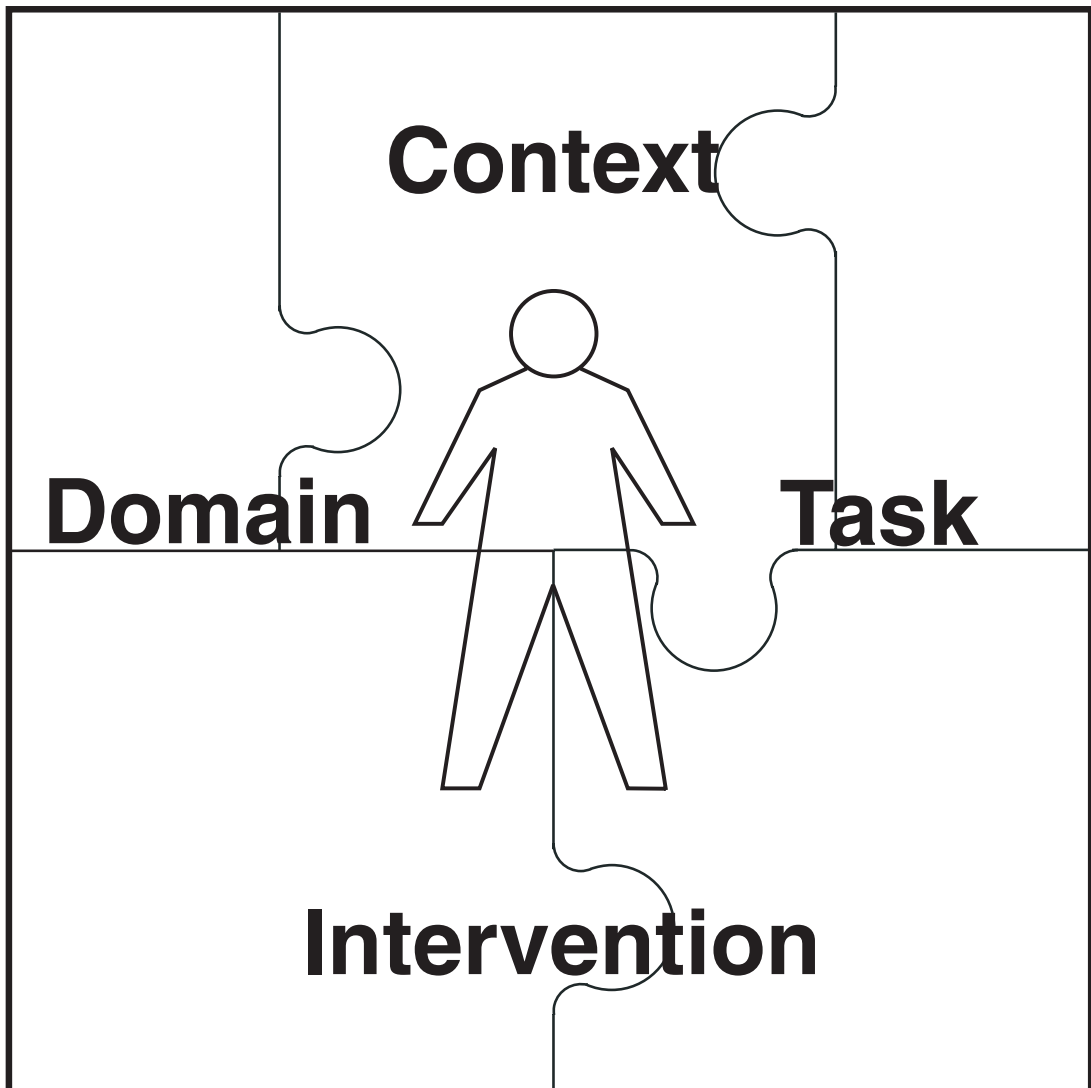
A framework that is particularly useful is one reliant on an ecological approach. From this perspective, it is critical to evaluate several different aspects of clients' lives in order to develop a better understanding of their functioning within a variety of contexts. The purpose of this chapter is to examine the particular areas of attention and memory, learning processes (input and output), and the special aptitudes of sensory perception, sensory-motor integration, and language/communication to facilitate making informed assessment decisions. One must possess knowledge of the cognitive processes that these tests purport to measure to make judgments about the usefulness of any given instrument with the client's presenting issue. Furthermore, the clinician is provided with an introduction to the types of

instruments used to measure memory, learning, and special aptitudes as well as a brief description of those instruments that have strong empirical support for use with children and adult populations. Any one of these areas, on its own, represents a very narrow picture of the overall functioning of an individual. However, when used in conjunction with a more thorough assessment, these areas can provide the missing pieces to the puzzle.

Client difficulties may be attributed not only to intra-individual characteristics, but also to the *domain* of functioning (e.g., social, vocational, educational); the *context* or environment in which the client is expected to function (e.g., job site, classroom, independent living); the requirements of particular assignments, jobs, or responsibilities, *task*; and the strategy used to teach or remediate a difficulty, *intervention* (Geil & D'Amato, 1996). Although the focus of this chapter is memory, learning, and special aptitudes, it may be helpful for the clinician to view these processes as key components within the conceptual framework presented in Figure 1. These areas represent extremely important aspects of psychological functioning and can help to complete the diagnostic picture of an individual by providing critical information to assist the clinician in accurate diagnosis and intervention planning.

4.09.1.1 Assessment Approaches

Before addressing the particular areas of concern, a brief discussion of the assessment process is warranted. Both quantitative and qualitative assessment procedures help to provide a breadth of information concerning individual functioning. A quantitative or product-oriented approach uses standard performance data to assess individuals within and across all the functional domains to be measured by comparing the findings to a normative group (D'Amato, Rothlisberg, & Rhodes, 1997; Dean, 1985a, 1985b; Lezak, 1995). This process detects whether the client's skills show a discrepancy when they are compared to other individuals performing within a normal range. Patterns of performance can also be carefully analyzed to determine the individual client's strengths and weaknesses. Data is usually considered in several ways: level of performance or current functioning (compared to normative standards); pattern of performance (uniqueness of strengths and weaknesses); right-left differences (comparing tests that evaluate both hemispheres including both sides of the body); pathognomic signs (indications of abnormal signs or brain damage); qualitative analysis (behavioral observations of problem solving); intervention



- Domain (social, vocational, educational)
- Context (job site, classroom, independent living)
- Task (assignment, job, responsibility)
- Intervention (remediation, counseling, training)

Figure 1 Conceptual framework of client functioning.

planning (recommendations for appropriate rehabilitation) (Hynd & Semrud-Clikeman, 1990; Jarvis & Barth, 1994; Reitan & Wolfson, 1985, 1993; Sattler, 1992; Selz, 1981).

Most proponents of a quantitative approach recommend a standard or fixed battery of tests. A fixed battery, such as a typical psychoeducational battery (i.e., Wechsler Adult Intelligence Scale-Revised [WAIS-R] or Wechsler Intelligence Scale for Children-3rd Edition [WISC-III], Minnesota Multiphasic Personality Inventory-2nd Edition, Bender Visual-Motor Gestalt Test, Woodcock-Johnson Psychoeducational Battery-Revised) or neuropsychologi-

cal battery (e.g., *Halstead-Reitan Neuropsychological Test Battery*; Reitan & Wolfson, 1993) involve the same set of instruments for each individual tested (Hynd & Semrud-Clikeman, 1990; Hynd & Willis, 1988). A standard battery format insures that a broad array of appropriate tools is used to cover all significant domains and therefore provide documented results that may be interpreted with ease. In fact, a standard battery approach may be the best choice when and if potential litigation is an issue because this method offers a normative data base to which client profiles can be compared and contrasted (Guilmette & Giuliano, 1991; Reitan & Wolfson,

1995). Despite the apparent strengths of a standardized approach, it has been argued that when developing treatment options, the use of qualitative methods that explore the process of learning or behavior may be better suited than a purely quantitative or product-oriented approach (D'Amato, Rothlisberg, & Leu, in press). For example, if verbal instruction with verbal response is a strength for the client, a preference for left hemisphere processing might be entertained and interventions utilizing a verbal component could be tailored with that hypothesis in mind. Likewise, if a client demonstrated a strength in simultaneous processing of information, a global concept or visual chart could be introduced before presenting the individual skills necessary to accomplish the particular task.

A second strategy, the qualitative approach or process-oriented approach, uses informal procedures such as direct observation of particular skills to analyze the specific patterns and processes in order to understand better the intricacies of the client's psychological processes (Lezak, 1995). Practitioners utilize a client's individual pattern of responses or results to guide the assessment process. That is, if a client was observed to have difficulty with memory tasks, that particular area would be investigated in more detail through the use of additional measures of memory. A decision-making process (i.e., whether to explore an area further or move on to another area of functioning) occurs after each item and is based on clinical judgment. By employing this strategy, it is argued that a clinician is better able to understand the complexities of an individual's performance and focus on the impaired functional system (D'Amato, Rothlisberg, & Leu, in press; Golden, 1981; Luria, 1980). Unique and individualized sets of procedures, questions, or tasks shape the evaluation process and might include an individual case study approach consisting of a mental status exam, observation, and symptom checklists. From an educational perspective, a psychologist using a qualitative approach might gather information using work samples, classroom observations, or dynamic assessment strategies (e.g., Campione & Brown, 1987; Feuerstein, Rand, & Hoffman, 1979). While the major areas traditionally covered in a qualitative evaluation seem comprehensive (e.g., investigations of motor functions, expressive speech, writing, reading; see Hynd & Semrud-Clikeman, 1990), this view does not rely on standardized batteries or clear comparisons to normative populations. Instead, the selection of strategies utilized follows significant clinical patient-practitioner interactions.

Although the flexibility and individualization apparent in this method is appealing, it requires a great deal of clinical experience to make accurate interpretations of behaviors, and problems with reliability and validity are ever-present (Lezak, 1995).

A third approach, and one which is likely used by the majority of clinicians, is the use of integrated data. Indeed, any time examiners note an examinee's reaction to a task, the response time involved, or any problem-solving strategies employed (e.g., rehearsal, verbal cuing), they are inferring the underlying processes being used (D'Amato, Rothlisberg, & Leu, in press; Taylor, 1988; Taylor & Fletcher, 1990). Because all individuals show a distinctive pattern of learning and behavioral characteristics, it is improbable that any given test, or even battery of tests, in isolation, can capture the range of skills exhibited by that individual. Furthermore, test scores that are interpreted without consideration to the context of the examination may be objective but are meaningless in their individual application (Lezak, 1995). Likewise, clinical observations unsupported by standardized and quantifiable testing may provide a rich picture of the client's current functioning but lack the comparability necessary for many diagnostic and planning decisions. Thus it is expected that most practitioners are integrated in their assessment practices, relying on both norm-referenced comparisons and qualitative procedures and observations to enrich their views of their clients. In fact, Lezak (1995) suggests that either method is incomplete without the other.

4.09.1.2 Evaluation of Domain Areas

Regardless of the position of the examiner along the quantitative-qualitative continuum, it is helpful to conceptualize an evaluation of domain areas, rather than simply focusing on tests or specific problem behaviors. The following domains are offered because of their importance to daily functioning and usefulness to intervention development in educational and vocational settings (Begali, 1994; D'Amato & Rothlisberg, 1992; D'Amato, Rothlisberg, & Leu, in press; Gaddes & Edgell, 1994). These domains include:

- (i) intelligence/cognitive abilities,
- (ii) personality/behavior/family information,
- (iii) memory and attention,
- (iv) learning processes,
- (v) academic achievement,
- (vi) sensory/perceptual systems,
- (vii) motor functions, and
- (viii) communication/language skills.

The areas of intelligence and personality assessment are covered in more depth in Chapters 8 and 12 of this volume. The remaining areas are divided into the subareas presented in Table 1 to provide a better understanding of the types of skills encompassed in each of these domains. Given the complexity of these areas, all should be considered both formally and informally. Direct observations and interviews with the client and family members are vital components in evaluating any individual's performance. The selection of tests utilized to evaluate these abilities will vary greatly depending on the unique needs of the individual, considered in tandem with the reason for referral.

Data on the functioning of these domains provides useful information for the clinical psychologist. As demonstrated in Table 2, a variety of testing instruments are appropriate in each of these areas. It should be noted that different authors have suggested various subsets of domains for analysis as well as recommending literally hundreds of other measures as appropriate for children and adults (Batchelor, 1996a; Begali, 1992; Dean & Gray, 1990; Gaddes & Edgell, 1994; Hynd & Willis, 1988; Lezak, 1995). Thus the instruments categorized in Table 2 represent only a sampling of available measures. The practitioner must take responsibility for carefully matching the individual with potential assessment options, after considering the distinct features of the instruments and the unique needs of the client.

4.09.1.3 Intervention Approaches

The referral question for any client is rarely "how is this individual functioning today?"; instead the referral source is most often interested in the extent of decline following an injury or illness, the expected future performance in school or work settings, or how to maximize a client's potential given certain difficulties (e.g., head injury, learning disability; Long, 1996). Several decisions must be made in relation to the intervention strategy and will be reliant on the quality of the information provided by the assessment. Intervention may be conceptualized using one of three approaches: remediation (retraining a previously learned skill), compensation (learning to use other strengths to offset a lost skill), or a combination of both (D'Amato & Rothlisberg, 1996). In particular, it is critical to determine the level of intervention on which to focus one's efforts and the ideal combination of strategies that will work best with an individual. Rehabilitative efforts emphasize enabling clients to reach their goals in educational, vocational,

Table 1 Subdomains of attention and memory, learning process, and special aptitudes.

<i>Attention and memory</i>
Attention
Concentration or vigilance
Visual memory
Verbal memory
Recall
Recognition
Short-term memory
Long-term memory
<i>Learning processes (input and output)</i>
Visual processing
Motoric processing
Auditory processing
Linguistic/verbal processing
Simultaneous processing
Sequential processing
Academic achievement
<i>Sensory/perceptual</i>
Visual
Auditory
Tactile-kinesthetic
Integrated
<i>Motor functions</i>
Strength
Speed
Coordination
Lateral preference
Sensory-motor integration
<i>Communication/language skills</i>
Receptive vocabulary
Expressive vocabulary
Speech/language
Written language

Source: Adapted from D'Amato & Rothlisberg (in press) and D'Amato, Rothlisberg & Rhodes (1997).

social, and recreational settings despite difficulties related to their deficits. To reach this end, intervention strategies may focus on:

- (i) remediating or retraining impaired cognitive processes (if there is a reason to believe that the process can be improved with practice),
- (ii) helping the client to develop new skills to compensate for residual deficits,
- (iii) creating classroom or workplace adaptations and other environmental compensations that permit effective performance despite residual deficits,
- (iv) choosing instructional or therapeutic procedures that best fit the client's profile of strengths and weaknesses, and
- (v) promoting improved metacognitive awareness of strengths and needs so that the

Table 2 Common instruments and procedures used to evaluate attention and memory, learning processes and special aptitudes.

Attention and memory

Test of Variables of Attention (TOVA™)
 Visual Search and Attention Test
 Tests of Memory and Learning
 Wechsler Memory Scale-Revised
 Wide range assessment of memory and learning

Learning processes: input and integration

Detroit Tests of Learning Aptitude-3
 Swanson's Cognitive Processing Test
 Children's Auditory Verbal Learning Test
 Tactile Performance Test (Halstead-Reitan Battery)
 Speech-Sounds Perception Test (Halstead-Reitan Battery)
 Wisconsin Card Sort

Academic achievement: output

Woodcock-Johnson Psycho-educational Battery-Revised: Achievement
 Peabody Individual Achievement Test-Revised
 Wechsler Individual Achievement Test
 Kaufman Test of Educational Achievement
 Keymath-revised
 Woodcock Reading Mastery Test-Revised
 Test of Reading Comprehension-3
 Test of Written Language-3

Sensory perception

Observations
 Developmental history
 Mental status examination
 Motor-free Visual Perception Test
 Vision and hearing screening

Motor (fine and gross)

Bender Visual-Motor Gestalt Test
 Detroit Test of Learning Aptitude-3 (Motoric Composite)
 Developmental Test of Visual-Motor Integration
 Finger Oscillation Test
 Grip Strength Test
 K-ABC Nonverbal Scale (e.g., Hand Movements subtest)
 McCarthy Scales of Children's Abilities (Motor Scale)
 WISC-III and WAIS-R (Block Design, Object Assembly, Coding subtests)
 Bruininks-Oseretsky Test of Motor Proficiency

Communication/language skills

Revised Token Test
 Peabody Picture Vocabulary Test-Revised
 Test of Adolescent Language
 Test of Language Development-2 (Primary and Intermediate)
 Test of Language Competence

client can become an active participant in selecting goals and interventions strategies (Ylvisaker, Szekeres, & Hartwick, 1994).

4.09.2 ASSESSMENT OF MEMORY

Memory is one of the most important cognitive functions to be assessed. It is a highly complex cognitive function that encompasses several relatively discrete stages: reception and registration of sensory stimuli, temporary short-term storage of information, storage of the information in a more permanent form (long-term memory), and recall and retrieval of previously stored information (Shiffrin & Atkinson, 1969; Taylor, Fletcher, & Satz, 1984). Functioning at each stage depends upon the integrity of the previous steps, with any interruption in the hierarchy having the potential to interfere with memory storage or retrieval. For example, difficulties with attention, which most closely relates to the first stage of memory, would obviously lead to problems in short- and long-term storage as well as later retrieval of the information. A further source of complexity in understanding and measuring this skill is the variety of theoretical approaches from which memory can be conceptualized, including information processing, neuropsychological, and behavioral perspectives.

Memory testing is very useful for assessing the possibility of organic disease, in helping to differentiate between organic and psychiatric disorders, and in determining the functional significance of a memory problem (Black & Strub, 1994). Most of the major neurobehavioral disorders such as dementia, confusional states, amnesia, material-specific memory/learning defects, and attentional dysfunction are those in which disturbances of memory and attention are the prominent clinical features (Hamsher, 1984). However, individuals with depression, post-traumatic stress disorder, dissociative disorders (e.g., dissociative amnesia, dissociative identity disorders) might also demonstrate attention and memory deficits (American Psychiatric Association, 1994). Memory skills represent a difficult area to address because of the variety of levels (e.g., working, short-term, long-term) and the potential implications of a deficit. So too, memory to some degree is modality specific; that is, for example, some individuals may have impaired verbal memory but intact visual memory. Thus, it is important to look at various components of memory rather than obtaining a simple global memory score. Tests that provide a single memory score offer a myopic and problematic view of the multifaceted quality of memory.

4.09.2.1 Attention

One of the key components to memory and learning is the ability to attend selectively to relevant information that we are presented with during the course of daily functioning. Attention is an extremely important basic function which refers to the client's ability to maintain awareness and to focus on a specified environmental stimulus, while screening out other stimuli that are potentially distracting (Black & Strub, 1994). Being able to attend has three major benefits for an individual: accuracy, speed, and maintenance of mental processing (LaBerge, 1995). Attention deficits appear as distractibility or impaired ability for focused behavior, regardless of the individual's intention (Lezak, 1995). Intact attention is a necessary condition of concentration which requires an individual to sustain attention over an extended period of time. Concentration problems may be due to a simple attentional disturbance, or to inability to maintain a purposeful attentional focus or, as is often the case, to both problems. This skill is important for adequate performance on any cognitive task, and can be impaired as a result of either an organic or emotional disorder (D'Amato, 1990; Dean, 1985a). Several psychological difficulties have been associated with attentional problems such as impulsivity, distractibility, and poor social judgment. Tests that require mental effort and persistence can measure an individual's ability to select, sustain, and shift attention (Slomka & Tarter, 1993). By comparing performance on various types of tasks, the practitioner is able to distinguish a global attention deficit from the more discrete, task-specific problems of concentration and tracking.

It is important to clarify the nature of an attention problem by observing people's general behavior as well as their performance on tests involving concentration. An interview with family members can provide important information about attentiveness and susceptibility to distraction. So too, formal or quantitative measures of attention and short-term memory can be derived from the Digit Span and Coding or Digit Symbol subtests of the WISC-III or WAIS-R tests. Some of the more specific, informal measures of attention and concentration might include observation, a digit span task, and a vigilance task as outlined by Strub and Black (1993). The individual is given orally a series of random letters with the letter "A" occurring with greater frequency than the other letters. The individual is instructed to signal whenever the targeted letter (i.e., A) is heard. The individual's performance is scored for

errors, which are rarely made by those without attention or vigilance difficulties (Black & Strub, 1994). More formal measures such as the *Visual Search and Attention Test* (Trenerry, Crosson, DeBoe, & Leber, 1990) can be used for adults. This test purports to measure sustained attention and visual scanning. The test consists of four 60-second trials and is made up of four tasks which become increasingly complex. The respondent is required to cross out letters or symbols that match a target. Normative tables are provided and arranged in four 10-year age bands, an 18–19 year age band, and a 60+ age band, and the statistical properties appear to be adequate (Hooper, 1995). Based on how the stimuli are presented, these tasks can provide a measure of either visual or auditory vigilance.

Vigilance tests are often referred to as continuous performance tests (CPTs), which are automated tasks, now computer-administered, that purport to measure sustained attention (Greenberg & Waldman, 1993; Lassiter, D'Amato, Raggio, Whitten, & Bardos, 1994; Rosvold, Mirsky, Sarason, Bransome, & Beck, 1956). CPTs have become a popular tool for clinicians to measure attentional performance, response inhibition, and medication monitoring in both children and adults (Eliason & Richman, 1987; Lassiter et al., 1994). Many versions of CPT have been developed since the original but the basic methodology of these tasks remains fairly constant. Clients are presented with a variety of stimuli that are displayed for a short period of time, and are instructed to respond to a predefined "target" stimulus. A number of different indices can be recorded with these tasks including omissions errors (i.e., failing to detect target stimulus), commissions (i.e., responding to nontarget stimulus), and response times for correct detections (Greenberg & Waldman, 1993). Commission errors are considered to be indicative of impulsivity and omission errors are thought to denote inattention (Eliason & Richman, 1987; Lassiter et al., 1994). Examples of these types of tests include the CPT-2 (Lindgren & Lyon, 1983), the *Raggio Evaluation of Attention Deficit Disorder* (Raggio, 1991), and the *Test of Variables of Attention (TOVA™)*; Greenberg, 1993).

The *TOVA™* is a nonlanguage-based, visual continuous performance test. This test runs for 23 minutes on a fixed-interval schedule and presents two easily discriminated visual stimuli for 100 milliseconds every two seconds. It was designed for use in the diagnosis and monitoring of pharmacotherapy of children and adults with attention deficit disorders and can be used with individuals age five to adulthood. The test does not require right-left discrimination and has

negligible practice effects. Recently, the authors of the *TOVA*TM have created developmental norms for children aged 6 to 16 which are available for few other CPT versions (Greenberg & Waldman, 1993). This type of tool may also be useful in assisting the clinician in the differential diagnosis of children and adolescents experiencing externalizing problems (e.g., attention deficit disorder, oppositional defiant disorder, conduct disorder, and aggression) and/or learning disabilities (Eliason & Richman, 1987; Greenberg & Waldman, 1993).

Despite the technological advances and newly defined norms, CPTs present a quandary to practitioners because of the variety of attributes that the tests reportedly measure. Some see these tests as measuring attention and impulsivity (Klee & Garfinkel, 1983), educational achievement (Campbell, D'Amato, Raggio, & Stephens, 1991), behavior (Lassiter et al. 1994), general neuropsychological functioning (Halperin, Sharma, Greenblatt, & Schwartz, 1991), and information processing (Swanson, 1981). Given this variance, a practitioner is left with the question of how to interpret the test results of a particular client. While research supports many of these claims, different versions of the CPTs have been used in these studies, with different samples of children and adults. So too, the validity of CPTs have been related to material collected from teachers, parents, and peers, and from standardized intelligence, achievement, and personality tests. While it is obvious that CPTs measure issues critical to learning and memory, the specificity of these instruments remains unclear. In conclusion, Morris (1996) noted that many of the measures that purported specifically to measure sustained attention often measured other variables, and thus many of these tests have poor construct validation and may be more appropriately viewed as multi-dimensional in nature. As an alternative, Barkley (1996) advocates the use of more natural tasks to study attention in an individual. He concludes that CPT-type tasks are unrelated to our daily functioning and thus, an individual's performance on such tasks is irrelevant. In response to this concern, several investigators have reportedly used television viewing, performance on classroom tasks, video games, and driving performance as a means of studying attention and its deficits in various groups of children and young adults (Barkley, 1996).

4.09.2.2 Short-term and Long-term Memory

As already noted, each component of the memory process is reliant upon the previous steps. If information in sensory storage undergoes additional processing, it becomes a more

lasting memory (short-term or long-term). Furthermore, these different memory functions must be systematically reviewed through visual and aural modalities using both recall and recognition tasks. Lezak (1995) suggests that at a minimum, the memory examination should include: *immediate retention* tasks, including short-term memory with interference; *learning* in terms of extent of recent memory, learning capacity, and how well newly learned material is retained; and efficiency of *retrieval* of both recently learned and long-stored information (i.e., remote memory).

Informal methods of assessment include tests of immediate recall such as digit repetition and/or sentence repetition, interviewing for information from remote memory (e.g., "where were you born?"), and new learning ability (e.g., immediate recall for a verbal story, asking the individual to remember four unrelated words for a span of 5, 10, and 30 minutes). During this last task, the examiner can provide recognition cues if the individual is having difficulty remembering the words. It is expected that those without difficulties will remember all words, while those with brain damage might be expected to remember one (Black & Strub, 1994). For aphasic clients or those with other speech or language problems, an informal measure of visual memory can be completed by hiding five objects around the interview room as the client names each item as it is hidden. After 10 minutes, the client is asked for name and location of each item. Reportedly, both normal and lower IQ clients should be able to find all five objects, with slightly lower performance for older patients (approximately four objects) (Black & Strub, 1994; Simpson, Black, & Strub, 1986). These memory tasks should be supplemented with observations and interviews with family members. So too, if an ability measure such as the WISC-III or WAIS-R is administered, performance on Digit Span can provide information on immediate verbal retention and the information subtest can be an indicator of the extent of remote memory in an individual.

To complete an assessment of the major dimensions of memory, Lezak (1995) has suggested including:

- (i) a test of configural recall and attention such as the visual reproduction subtest on the *Wechsler Memory Scale* (Wechsler, 1987) or the *Benton Visual Retention Test* (Benton-Sivan, 1992);
- (ii) a paragraph for recall to examine learning and retention of meaningful verbal material; and
- (iii) a test of learning ability that gives a learning curve and includes a recognition trial,

such as Rey's Auditory-Verbal Learning Test (for review see Lezak, 1995).

These techniques should be integrated into the general clinical interview to create a varied testing format, to enable the practitioner to use nonmemory tasks as interference activities, and to reduce stress in those clients who have memory impairments and are concerned about their abilities (Black & Strub, 1994).

There are numerous formal instruments available which measure different dimensions of memory. For children and adolescents, the *Wide Range Assessment of Memory and Learning* (Sheslow & Adams, 1990), and the *Test of Memory and Learning (TOMAL)* (Reynolds & Bigler, 1994) can be used to evaluate individual strengths and weaknesses in the areas of memory and attention. In particular, the *TOMAL* represents a reliable, empirically sound measure for children and adolescents. The *TOMAL* consists of four core indexes comprising Verbal Memory, Nonverbal Memory, Composite Memory, and Delayed Recall. Supplementary indexes for Learning, Attention and Concentration, Sequential Memory, Free Recall, and Associative Recall are also provided. Subtests include Memory for Stories, Facial Memory, Word Selective Reminding, Visual Selective Reminding, Object Recall, Abstract Visual Memory, Digits Forward, Visual Sequential Memory, Paired Recall, Memory-for-Location, Manual Imitation, Letters Forward, Digits Backward, and Letters Backward. The *TOMAL* was standardized for children aged 5 to 19.

The *TOMAL* boasts many unique features, including a great variety of memory indexes (Reynolds & Bigler, 1994). While some of the subtests appear similar to other memory measures, some unique features of this test include a learning index where teaching is permissible, a sequential memory index, and an attention and concentration index. Delayed recall subtests are also available and are offered as an evaluation of forgetting or memory decay. It is possible to compare the examinee's own personal learning curve with a standardized learning curve. The test is easy to administer and generally user-friendly. Its psychometric properties appear to be well-developed. In the *TOMAL* subtests, 63% of the reliability coefficients are at or exceed 0.9, 31% are between 0.8 and 0.89, and only 6% fall below 0.8. Test-retest coefficients range from 0.71 to 0.91. Support for the validity of this instrument was determined through indices of content validity, construct validity (e.g., factor analytic studies), and criterion-related validity.

Assessment of memory dysfunction in adults is easier than in children because their period of

rapid intellectual, academic, and physical (including neurological) development has ended (Reynolds & Bigler, 1994). In adults, memory dysfunction is associated with a variety of well-defined disorders, and in many individuals is one of the earliest and key symptoms such as in Korsakoff's disease and various other dementias including Alzheimer's disease. Because of the key role of evaluating memory in the clinical setting, there are a number of instruments designed for memory assessment in older populations including the *Doors and People: A Test of Visual and Verbal Recall and Recognition* (Baddeley, Emslie, & Nimmo-Smith, 1994), the *Memory Assessment Scales (MAS)* (Williams, 1991), and the *Wechsler Memory Scale-Revised (WMS-R)* (Wechsler, 1987).

The *WMS-R* (Wechsler, 1987) provides an extensive measure of several dimensions of memory. It consists of eight short-term memory tests, four delayed-recall subtests, and a brief screening measure of mental status (i.e., information and orientation questions). The eight short-term memory tests yield four composite scores: Verbal Memory, Visual Memory, Total General Memory, and Attention/Concentration. The delayed-recall measures can be combined to derive a fifth composite score, Delayed Recall. The test is intended for use for individuals ranging in age from 16 to 74 and requires approximately 50 minutes to administer. The psychometric properties of the *WMS-R* are questionable in terms of low reliability coefficients for the composite scores (average $r = 0.74$), but provides stronger support for the General Memory and Attention/Concentration (average $r = 0.81$) scores. Although the *WMS-R* demonstrated satisfactory discrimination power between various clinical groups, factor analyses supported a two-factor rather than the hypothesized five-factor model. Huebner (1992) concluded that this instrument must be used cautiously in making clinical decisions about individuals and interpretation should be restricted to General Memory and Attention/Concentration ability.

For adolescents and adults, the *MAS* (Williams, 1991) also provides a valid, reliable, and comprehensive measure of memory functioning. The *MAS* was standardized for use with adults aged 18 to 90. The major functions measured by the *MAS* include: verbal and nonverbal learning and immediate memory; verbal and nonverbal attention, concentration, and short-term memory; and memory for verbal and nonverbal material following delay. In addition, measures of recognition, intrusions during verbal learning recall, and retrieval strategies are also available. The test consists of 12 subtests based on seven memory tasks.

Five of the subtests assess the retention of information learned in a subtest administered earlier in the sequence. Total testing time is approximately one hour. Test-retest reliability for the *MAS* was estimated using generalizability coefficients and these correlations averaged 0.85 for the subtests, 0.9 for the summary scales (i.e., Short-Term Memory, Verbal Memory, and Visual Memory), and 0.95 for the global memory scale. The validity of the *MAS* was established using three types of studies: convergent and discriminant validity, factorial validity, and group differentiation. Despite these strengths, Berk (1995) concluded that clinicians should use caution in interpreting the scores until some technical problems (e.g., inadequate samples, lack of evidence for content validity) can be corrected.

4.09.2.3 Memory: Implications for Intervention

Because memory is multifaceted, interventions in the memory domain must also be multidimensional. Interventions may be divided several ways; those involving language, those that are nonverbal, those requiring long, short, or intermediate memory, and those that use a combined approach to aid in retention (Gaddes & Edgell, 1994; Lezak, 1995). Strategy selection depends on accessing the strengths of the clients or, in the case of injury or disease, accessing those parts of the brain which have been least impacted. For example, learners who have difficulty with nonverbal memory tasks but have retained verbal skills may benefit from memory interventions that use language. Mnemonic devices may be used to assist in recall of information if a series of problem-solving steps is required (Mastropieri & Scruggs, 1989). For those with more difficulty remembering, some simple techniques such as writing all meetings in an appointment book or using grocery shopping lists and daily "to do" lists are practical.

4.09.3 ASSESSMENT OF LEARNING

Memory is a ubiquitous component of daily life and is fundamental to the process of learning. One must be able to remember in order to demonstrate learning. The classic definition of learning describes it as changes in behavior as a result of experience. Some have even considered this definition of learning as also defining memory (e.g., see Kolb & Whishaw, 1990). Despite this relatively simple definition, the learning process itself defies easy explanation. Learning can be approached from a neuropsychological (e.g., planning, attention, simultaneous, successive [PASS] model; Das, Naglieri, & Kirby, 1994), behavioral (e.g.,

classical and operant conditioning; Baldwin & Baldwin, 1986), cognitive (e.g., information processing; Pressley & Levin, 1983), or social (e.g., social learning and modeling; Bandura, 1977) perspectives, to name but a few. Furthermore, the distinction drawn between measures of memory and measures of learning is tenuous at best since all instruments that evaluate an individual's learning process will automatically include aspects of memory functioning. Because learning takes center stage as one of our functions of daily living, it represents a critical area to evaluate. By fully evaluating several components of learning, clinicians can determine where the process is breaking down and provide recommendations for rehabilitation. An understanding of how a client best learns also has important implications for the types of therapeutic strategies that will most likely be successful. For example, if an adolescent has difficulty processing and remembering auditory information, talk therapy may not be the best approach. Supplementing discussion with role play, videos, and other visual cues may be necessary to facilitate the client's acquisition of new knowledge.

Difficulties in learning can be attributed to a number of disorders, including the general category of learning disorders (e.g., dyslexia, dyscalculia), traumatic brain injuries, drug and alcohol abuse, and medical disorders (e.g., strokes, Alzheimer's disease). Indeed, all of the variables that can affect attention and memory will also impact learning. Furthermore, certain chronic medical conditions can play a role in learning difficulties. For example, childhood diabetes is associated with subtle problems with respect to visuospatial and visuomotor processing (e.g., Rovet, Ehrlich, & Hoppe, 1988), verbal abilities (Kovacs, Goldston, & Ivengar, 1992), and memory and attention problems, which translate into increased risk for difficulties in academic achievement (Kovacs et al., 1992; Rovet, Ehrlich, Czuchta & Akler, 1993). So too, sickle cell anemia, an inherited disorder in people of African descent, often produces some subtle cognitive impairments that can affect school achievement negatively (Brown, Armstrong, & Eckman, 1993). Among individuals with traumatic brain injuries, learning process problems may be reflected as an uncertainty as to whether a concept has been learned or not (Cohen, 1991).

4.09.3.1 Models of Learning

Information-processing theories have proved extremely useful in conceptualizing learning because this model can be applied to any given cognitive task and allows the practitioner to specify where the learning process is breaking

down. Silver (1993) proposed an information-processing model based on four steps: input (how information from the sense organs enters the brain), integration (interpreting and processing the information), storage (storing the information for later retrieval), and output (expressing information via language or muscle activity). Learning is reliant upon each of the first three steps and is observed or inferred from the fourth step. Other models of information processing highlight the importance of the working memory in skill acquisition and learning (Baddeley, 1986; Just & Carpenter, 1992; Swanson, 1995). Working memory has traditionally been defined as a system of limited capacity for the temporary maintenance and manipulation of information (e.g., Baddeley, 1986; Just & Carpenter, 1992) and most closely corresponds to the integration step in Silver's model. Tasks that measure working memory are those that require the client to remember a small amount of material for a short time while simultaneously carrying out further operations. In daily life, these tasks might include remembering a person's address while listening to instructions about how to reach a specific destination (Swanson, 1995). When viewed from this perspective, working memory differs from the related concept of short-term memory which is typically described as remembering small amounts of material and reproducing it without integrating or transforming the information in any way (e.g., repeating back a series of numbers) (Cantor, Engle, & Hamilton, 1991; Just & Carpenter, 1992). Working memory appears to be extremely important to an individual's ability to learn, and in adult samples has correlations of 0.55–0.92 with reading and intelligence measures (e.g., Daneman & Carpenter, 1980; Kyllonen & Christal, 1990).

In an effort to promote the notion that input and integration of stimuli can impact subsequent learning, Cronbach and Snow (1977) have advanced a theory suggesting that some types of individuals might benefit from one form of treatment, whereas others might benefit from another type of treatment: an aptitude by treatment interaction (ATI). Many researchers and educators alike believe that matching learner characteristics with treatment approaches can enhance learning (e.g., Cronbach & Snow, 1977; Resnick, 1976; Reynolds, 1981b). However, subsequent studies have demonstrated little support for this theory (e.g., Arter & Jenkins, 1977; Tarver & Dawson, 1978). Initially, theories of input examined learner modalities (e.g., visual, auditory, kinesthetic), which were later deemed to be too simplistic (Arter & Jenkins, 1977; Kaufman, 1994; Tarver

& Dawson, 1978). More recently, neuropsychological models have been applied to ATIs and offer promise for identifying aptitudes and prescribing treatments (D'Amato, 1990; Hartlage & Telzrow, 1983). One of the major techniques that Cronbach and Snow (1977) suggested for matching treatment approaches with learner aptitudes was "capitalization of strengths." Our increasing knowledge of how the brain functions allows clinicians to obtain a more detailed understanding of how a client learns new information. For example, although the cerebral hemispheres act in concert, the right hemisphere seems to be specialized for holistic, spatial, and/or nonverbal reasoning whereas the left shows a preference for verbal, serial, and/or analytic type tasks (Gaddes & Edgell, 1994; Lezak, 1995; Reynolds, 1981a; Walsh, 1978). Similarly, models of cognitive processing have been proposed that agree with the specialization of how scientists think the brain processes information; some have called this preferential processing styles (D'Amato, 1990). For example, simultaneous processing ability has been affiliated with the right hemisphere because of its holistic nature; it deals with the synthesis of parts into wholes and is often implicitly spatial (Das, Kirby, & Jarman, 1979). In contrast, the left hemisphere processes information using a more successive/sequential method, considering serial or temporal order of input (Dean, 1984, 1986). Models of brain organization have also been proposed that attempt to explain the diversity and complexity of behavior.

An expansion of the hemispheric specialization approach is offered in the planning, attention, simultaneous, successive (PASS) cognitive processing model (Das et al., 1994) which proposes four processing components. This model is based on the neuropsychological model of Luria (1970, 1973, 1980; Reynolds, 1981a) and presents a comprehensive theoretical model by which cognitive processes can be examined. On the basis of his clinical investigations with brain-injured patients, Luria (1973) suggested that there are three functional units that provide three classes of cognitive processes (i.e., memory, conceptual, and perceptual) responsible for all mental activity. Figure 2 provides a graphic presentation of the PASS model of cognitive processing. The functional units work in concert to produce behavior and provide arousal and attentional (first unit), simultaneous-successive (second unit), and planning (third unit) cognitive processes. The PASS model separates the second unit into two individual processes (i.e., simultaneous and sequential). Instruments can be used to measure individual strengths in these different styles of processing.

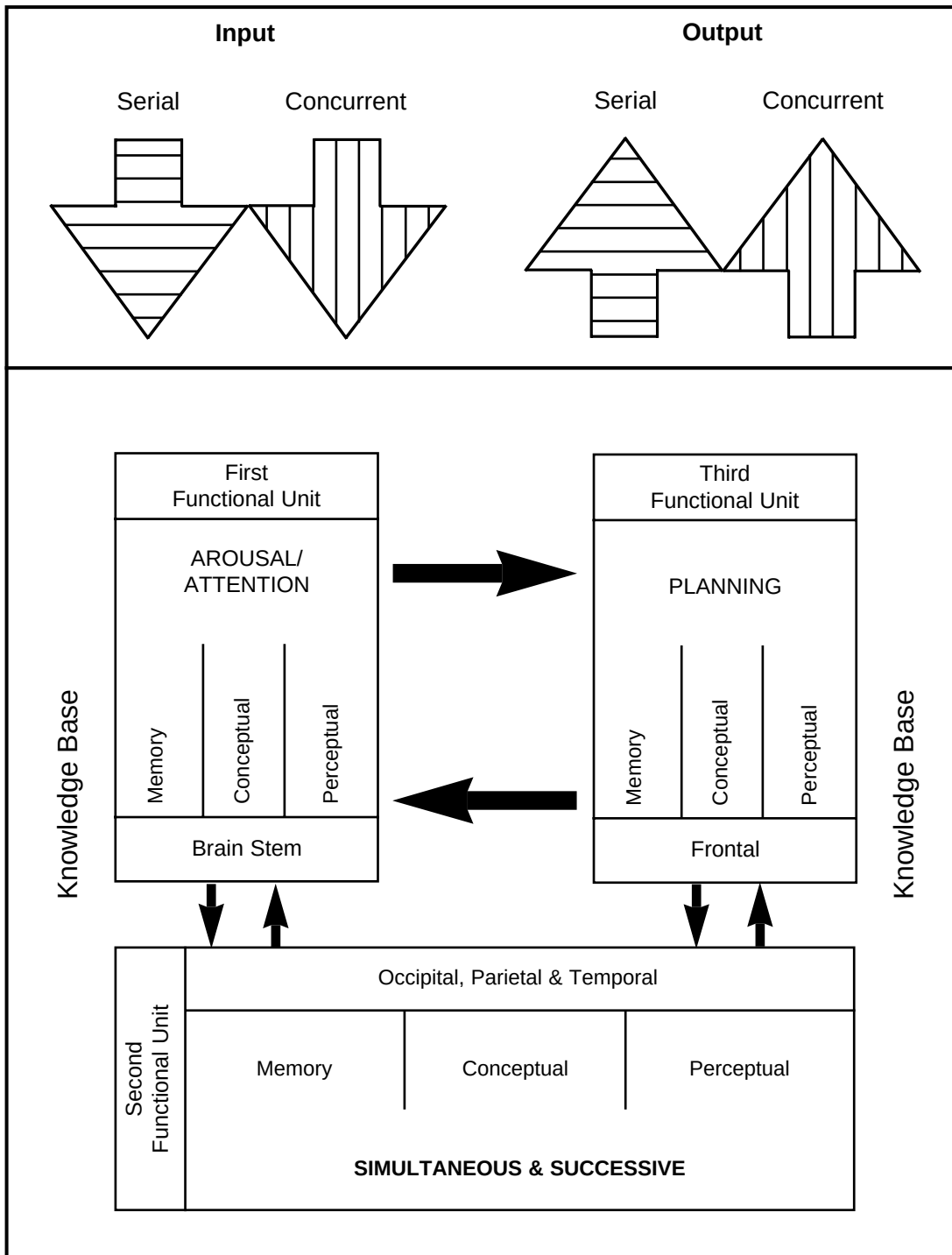


Figure 2 PASS model of cognitive processing. (*Assessment of Cognitive Processes: The Pass Theory of Intelligence* (p. 21), by J. P. Das, J. A. Naglieri, and J. R. Kirby, 1994, New York: Allyn & Bacon. Copyright 1994, by Allyn & Bacon. Reprinted with permission.)

Knowledge of the brain and theories governing information processing can determine the types of data collected during the assessment phase. For example, instead of simply observing whether the individual was successful at a task or set of measures, the practitioner looks

beyond the product to determine the influence of related factors. These factors can include the nature of the stimuli used (visual, verbal, tactile), the method of presentation (visual, verbal, concrete, social), the type of response desired (verbal, motor, constructional), and the

response time allowed (timed, untimed; Cooley & Morris, 1990). Other researchers have advocated a move to an even more intense examination of processing through the use of dynamic assessment strategies (Campioné & Brown, 1987; Feuerstein et al., 1979; Palincsar, Brown, & Campione, 1991). Theoretically, this strategy allows the examiner to obtain information about the client's responsiveness to hints or probes, and thus elicits processing potential (Swanson, 1995). When an examinee is having difficulty, the examiner attempts to move the individual from failure to success by modifying the format, providing more trials, providing information on successful strategies, or offering increasingly more direct cues, hints, or prompts (Swanson, 1995). This approach allows the examiner an opportunity to evaluate performance change in the examinee with and without assistance. However, there is little if any standardized information available on this technique and it has been criticized for its clinical nature and poor reliability (e.g., Palincsar et al., 1991).

4.09.3.2 Learning Processes: Input and Integration

Generally, three approaches have been utilized when evaluating how individuals preferentially process information. The first approach, seen as the traditional approach, employs established measures (such as the WISC-III) with the practitioner seeking to understand information processing through an analysis of common test results such as reviewing global scores, subtests, and clusters of subtests (Kaufman, 1990, 1994). For example, a pattern of strengths on the Picture Completion, Block Design, Object Assembly and corresponding weaknesses in Picture Arrangement and Coding might suggest meaningful differences in a client's mental processing style (i.e., right hemispheric functioning vs. left hemispheric functioning from cerebral specialization theory, or simultaneous vs. successive coding from Luria theory). The second view of information processing, considered the informal approach, considers observations, checklists, and learning style inventories to understand how individuals learn. From this view, individuals who seem to profit most from visual clues may be seen as visual learners, and might be taught utilizing overheads, visual diagrams, and worksheets. The final approach to understanding processing stems from the administration and analysis of the many unique measures that have been offered as learning style or processing tests. This approach is seen as a nontraditional test approach (D'Amato, Roth-

lisberg, & Rhodes, 1997). These specialized measures of performance in learning do not fall neatly within the traditional domains of intelligence, achievement, or neuropsychological processing. These tests, including the *Detroit Tests of Learning Aptitude-3* (DTLA-3; Hammill, 1991), the *Swanson Cognitive Processing Test* (S-CPT; Swanson, 1996), the *Children's Auditory Verbal Learning Test-2* (Talley, 1993), and others can offer valuable information concerning how individuals attend to and deal with new information. While practitioners have used these instruments to document clients' strengths and weaknesses, diagnose problems, and chart the course of disorders, these instruments offer more practical information concerning rehabilitation or program planning than for diagnostic activities.

Neuropsychological tests have also been seen by some to evaluate variables related to learning processes. In fact, the *Halstead Reitan Neuropsychological Test Battery* (Reitan & Wolfson, 1985, 1993) reportedly measures problem solving, tactual discrimination, sensory recognition, spatial memory, verbal-auditory discrimination, attention, nonverbal auditory discrimination, psychomotor speed, and manual dexterity as well as several other skills (D'Amato, 1990; Dean, 1985a, 1985b, 1986; Lezak, 1995). Individuals interested in a neuropsychological approach to processing should consult some of the recommended references and Chapters 10 and 11, this volume.

The area of learning processing is more difficult to evaluate and is often subsumed within the intelligence or achievement domains. One measure that has a long history in the evaluation of processing styles is the *DTLA-3*. This instrument was designed for use with individuals aged 6 to 17. More recently, two other versions of this test, the *Detroit Tests of Learning Aptitude-Adult* (Hammill & Bryant, 1991a) and *Detroit Test of Learning Aptitude-Primary* (2nd ed.; Hammill & Bryant, 1991b) have expanded the usefulness of this instrument to include individuals from age 2 to 79. The *DTLA-3* consists of 11 subtests comprising: Word Opposites, Design Sequences, Sentence Imitations, Reversed Letters, Story Construction, Design Reproduction, Basic Information, Symbolic Relations, Word Sequences, Story Sequences, Picture Fragments; and 16 composite scores: General Mental Ability Composite, Optimal Level Composite, Domain Composites (Verbal, Nonverbal, Attention-Enhanced, Attention-Reduced, Motor-Enhanced, Motor-Reduced), Theoretical Composites (Fluid Intelligence, Crystallized Intelligence, Associative Level, Cognitive Level, Simultaneous Processing, Successive Processing, Verbal Scale, and

Performance Scale). The testing time is estimated to vary from 50 minutes to two hours. The internal consistency reliabilities for the subtests are sufficiently high; however, the data on stability are limited. It was also noted that the factor analysis does not support the construct validity of the different composites. In fact, only four factors (one being a residual or difficult to interpret category) were identified in the manual. Despite these concerns, Poteat (1995) notes that the *DTLA-3* can be recommended as an adjunct to some of the better developed measures of intelligence and it provides some potentially valuable information about diverse abilities. The *DTLA* has been especially helpful when evaluating children who suffer from learning disabilities or traumatic brain injuries. The *Detroit Tests of Learning Aptitude-Adult* (Hammill & Bryant, 1991a) comprises 12 subtests and 16 composites and measures areas similar to the *DTLA-3*. Internal consistency reliability of all scores approximates 0.9 for all ages. This instrument represents a useful tool in practice because of the type of information it can provide regarding client cognitive functioning in relation to learning new information.

A very recent contribution to the field of cognitive processing is the *S-CPT* (Swanson, 1996) which purports to measure different aspects of intellectual functioning and information processing potential. The battery, designed for use with persons age five to adulthood, draws from the work on information processing theory and dynamic assessment. The subtests in this measure are as follows: Rhyming Words, Visual Matrix, Auditory Digit Sequencing, Mapping and Directions, Story Retelling, Picture Sequencing, Phrase Recall, Spatial Organization, Semantic Association, Semantic Categorization, and Nonverbal Sequencing. This standardized test battery can be administered in an abbreviated form (five subtests) or in a complete form under traditional or interactive testing conditions. Normative data for the *S-CPT* were gathered on 1611 children and adults. The author reports high levels of internal reliability and high construct and criterion-related validity (Swanson, 1995). This instrument may offer a promising alternative to product-oriented evaluation strategies while still allowing for normative comparison.

4.09.3.3 Academic Achievement: Output

It is likely that those individuals who experience difficulties in processing or learning will display academic difficulties as well. Although some might hold that there is little difference in the measure of ability and the

measure of achievement (e.g., Anastasi, 1988; Dean, 1977, 1983), it would seem that the operationalization of the two areas allows for a comparison of more generic problem-solving and verbal tasks to those directly involved in scholastic performance. Thus, a measure of ability may be conceived of as attempting to address the concept of underlying skills or capacities, whereas the measure of achievement is tied to the notion of the individual's proficiency in applying that ability in a functional way to real world skills (e.g., academics). A measure of academic achievement can help provide information as to the degree of impairment experienced by individuals, especially among children and adolescents. New learning, however, is not isolated to the school years. Adults required to learn new skills as part of job training, vocational rehabilitation, or after brain injuries are all placed in very real learning situations. It is critical to have an understanding of the client's basic skills in order to facilitate vocational, academic, and intervention decision making.

Assessment of academic achievement can occur through a blend of formal and informal measures as well. For example, in a school setting, reviewing student clients' work samples, interviewing the students and teacher about their learning and the classroom, and classroom observations can provide essential information. So too, curriculum-based measurement, where informal reading, writing, and math probes (Shinn, 1989) are obtained to determine the clients' current level of functioning and progress during intervention phases, are particularly useful for monitoring the effectiveness of treatment approaches to learning difficulties (Fuchs, 1994). There are also several types of norm-referenced instruments that are available, which, because of the availability of a standard normative base, permit comparison across a wide variety of curricular contexts (D'Amato, Rothlisberg, & Rhodes, 1997). Some of these instruments measure a particular area of achievement such as math or reading (e.g., *Keymath Revised*, Connolly, 1988; *Test of Reading Comprehension-3*, Brown, Hammill, & Wiederholt, 1995; *Test of Written Language-3*, Hammill & Larsen, 1996) while others provide a broad-based screening of a number of academic areas (e.g., *Peabody Individual Achievement Test-Revised [PIAT-R]*, Markwardt, 1989; *Woodcock-Johnson Psychoeducational Battery-Revised [WJPB-R]*, Woodcock & Johnson, 1989). These broad-based tests all have a similar organizational structure. For example, measures in a particular area, such as reading, are typically divided into basic skill areas (e.g., reading decoding) and some form of

applied skill area (e.g., reading comprehension) so that variations in the aspects of the academic tasks can be noted. The difference between measures often lies in the method by which they obtain their information (e.g., whether visual-motor or oral responses are required); that is, whether they require the examinee to indicate the response through nonverbal (e.g., pointing) or verbal output.

The achievement test that is designed for the broadest range of individuals is the *WJPB-R* with norms ranging from 2 to 95 years of age. The *WJPB-R* consists of both a cognitive and an achievement component. The tests of achievement are divided into a standard battery consisting of four broad areas: Reading (Letter-Word Identification, Passage Comprehension), Mathematics (Calculations, Applied Problems), Written Language (Dictation, Writing Samples), and Broad Knowledge (Science, Social Studies, Humanities). A supplemental battery is also available to expand the standard battery coverage. It includes Word Attack, Reading Vocabulary, Quantitative Concepts, Proofing, and Writing Fluency. Employing one or more of the supplemental subtests gives the examiner the option of computing additional areas of achievement such as Basic Reading Skills and Reading Comprehension which is consistent with the language of the Individuals with Disabilities Educational Act of 1990 and some state legislative guidelines for identifying specific areas of learning disability. This instrument is statistically sound and ample amounts of research have been conducted and support the use of this test.

Another measure of achievement, the *PIAT-R* (Markwardt, 1989) has also been supported as a well-developed and psychometrically sound instrument (Williams & Vincent, 1991). It consists of five subtest scores (General Information, Reading Recognition, Reading Comprehension, Mathematics, Spelling) that are provided in addition to the Total Reading and Total Test scores. A Written Expression and optional Written Language score are also available. The *PIAT-R* was normed for individuals aged 5–18 years. It is different from other tests in that it includes a larger pictorial component in its item types, letting children avoid the need for verbal reply, and instead expecting them to point at the correct answer (out of four) for reading, spelling, and mathematics items. Since the task demands for recognition of information do not appear to be the same as for recall, this response format may aid children with retrieval difficulties or those that have developed some background knowledge of the area in question. It should be noted, though, that this response-type advan-

tage may not give a good indication of the expectations for student performance in the classroom where recall and more integrated answers are the norm (D'Amato, Rothlisberg, & Rhodes, 1997).

4.09.3.4 Learning: Implications for Intervention

A number of authors have related how knowledge of the way individuals process information can contribute to the development of treatment based on neuropsychological processes (D'Amato, 1990; Reynolds, 1981b, 1986; Telzrow, 1985). For example, when learning how to read, individuals who display a simultaneous/visual spatial strength in processing might benefit from being taught using a whole word approach whereas individuals who display a strength in sequential/auditory processing can be taught using a phonetic approach (Whitten, D'Amato, & Chittooran, 1992). For both children and adults, cognitive rehabilitation is an emerging discipline which includes the retraining or use of compensatory strategies in thinking and problem-solving skills (Wedding, Horton, & Webster, 1986). Cognitive retraining can include assistance in strategy development for attention and concentration, memory, language, perceptual and cognitive deficits, and social behavior. Thus, the term cognitive retraining encompasses all areas of functioning that may have been negatively impacted by neuropsychological disorders or traumatic brain injury (D'Amato, Rothlisberg, & Leu, in press; Gray & Dean, 1989). Assisting learners with cognitive remediation or compensation often includes the use of metacognitive strategies. Metacognition includes analyzing the processes an individual uses to generate an idea or thought. By receiving assistance in breaking down problems and understanding the processes needed to solve problems, clients may learn how to generalize the process to many problem types and improve overall learning and functioning. Although cognitive retraining is time consuming, the generalizability of the strategies has been seen as appropriate to many settings (Gray & Dean, 1989; Kavale, Forness, & Bender, 1988).

4.09.4 ASSESSMENT OF SPECIAL APTITUDES

In determining the basis for a client's difficulty, it is critical to explore the building blocks of memory and learning to obtain an understanding of how the individual processes information (sensory input). For instance, sensory and perceptual skills are essential to

receiving stimuli from the environment and making sense of what is received. So too, a clinician must examine the output or production that the client demonstrates in response to stimuli via action (e.g., motor skills) or communication (e.g., spoken language, writing). That is, clients may understand a task, but, because of integration difficulties or language impairments, be unable to demonstrate their knowledge. For example, the reproduction of a visual stimulus in response to a request involves both perceptual discrimination and fine motor development, as well as the ability to integrate visual, tactile, and auditory skills. Therefore, inadequate performance in copying geometric designs developed to assess these skills may stem from: a misperception, or faulty interpretation of the input information; problems in executing the fine motor response, or output; and/or difficulties integrating the input and output, otherwise known as integrative or central processing difficulties. By evaluating the domains of sensory perception, sensory-motor integration, and communication/language, the practitioner is in a better position to understand the client's ability to receive information adequately, integrate these basic skills, and demonstrate the products of memory and learning processes.

4.09.4.1 Sensory Perception

Perception of stimuli is a complex process involving many different aspects of brain functioning (Lezak, 1983). Typically, perception includes recognizing features and relationships among features. It is affected by context (figure-ground) and intensity, duration, significance, and familiarity of the stimuli (Ylvisaker, Szekeres, & Hartwick, 1994). Sensory perception skills are vital to an individual's understanding and response to the environment because they form the basis of each individual's interaction with the world (D'Amato, Rothlisberg, & Rhodes, 1997; Lezak, 1995). Difficulties may manifest themselves in the individual's ability to use information gained through the senses. For example, a client may be able to hear sounds well, but have trouble understanding what is heard (auditory processing). Likewise, a client may be able to see words clearly but have problems reproducing them when writing (visual-motor difficulties). Sensory perception tasks often form the foundation for the later performance of higher order cognitive skills. Without the ability to accurately sense and perceive cues from the environment, the learner is placed in the position of trying to decode a message when the code is scrambled and often changing.

In the assessment of sensory perception, it is important to evaluate visual, auditory, and haptic (tactile) functions. For children and older adults it is especially important that actual sensory deficits have been ruled out through the administration of a thorough vision and hearing test. If these senses appear to be intact, an in-depth evaluation of functioning is warranted. Sensory perception can be evaluated informally through clinical observations, formally through standardized tests, or via other methods of data collection. However, at times these strategies may prove inconclusive regarding the etiology of performance difficulties and more formal assessment is necessary to evaluate a client's functioning.

Several instruments are available to measure a client's functioning within this domain. Most are inexpensive and relatively quick to administer. Within the visual modality, the *Motor-Free Visual Perception Test (MFVPT)* (Colarusso & Hammill, 1996) allows the clinician to evaluate visual perception without motor involvement in children. This 36-item measure assesses five facets of visual perception: spatial relations, visual discrimination, figure-ground, visual closure, and visual memory. The *MFVPT* is intended for children four to eight years of age. The *MFVPT* can offer information essential for the differential diagnosis of motor vs. visual processing problems. However, when used in isolation from other measures or techniques, the *MFVPT* offers information regarding visual processing difficulties but is unable to rule out motor concerns. For adults, the *Benton Revised Visual Retention Test* (Benton-Sivan, 1992) is a widely used measure of visuo-perceptual ability, constructional skills, and immediate visual memory (Youngjohn, Larrabee, & Crook, 1993). Clients are required to reproduce abstract geometric designs from memory.

Some clients have difficulty discriminating sounds even when thresholds for sound perceptions are intact (Lezak, 1995). Auditory discrimination can be tested by having the client repeat words and phrases spoken by the clinician, or by asking the client whether two spoken words are the same or different. On this task, the clinician will want to use word pairs that sound alike such as "cat" and "cap" along with identical word pairs (Lezak, 1995). This technique has been formalized through the development of *Wepman's Auditory Discrimination Test* (Wepman & Reynolds, 1987) which allows the clinician to determine whether the client is able to discriminate similar sounding words adequately. Although the test was originally devised to identify auditory discrimination problems in young school children,

and the present norms were developed on samples of four to eight year olds, norms for the oldest age group (8–0 to 8–11) are adequate for adults since auditory discrimination is generally fully developed by this age (Lezak, 1995).

The perception of tactile stimuli is regularly measured as a component of a thorough neuropsychological examination, but less often in nonspecialized clinical settings. Informal strategies for evaluating this area include asking clients to indicate whether they feel the sharp or the dull end of a pin, pressure from one or two points (applied simultaneously and close together), or pressure from a graded set of plastic hairs, the “Von Frey hairs” (Lezak, 1995). The eyes should be closed or the hand kept out of sight when tactile sensory functions are tested. More formal measures include the *Tactile Form Perception Test* (Benton, Hamsher, Varney, & Spreen, 1983) and the *Tactual Performance Test* (Reitan & Davison, 1974). Deficits in tactile senses are often associated with damage to the right hemisphere of the brain and may have important implications for a client’s vocational functioning (Lezak, 1995).

4.09.4.1.1 Sensory perception: implications for intervention

If the client is having difficulty in one or more areas of sensory perception, this information is critical for intervention planning. That is, the client’s unique pattern of receiving information from the environment can be used to create effective education, rehabilitation, or therapeutic intervention. If a client is weak in auditory processing but strong in visual processing, for example, visual cues such as drawings, videos, or demonstrations may be the most effective means for training them in new skills.

4.09.4.2 Motor: Fine and Gross

The motor domain involves a range of both fine and gross motor movement. Fine motor skill is commonly thought of as movement which does not involve the entire body. Writing, opening a letter, or tying a shoe are all examples of fine motor movements. Gross motor movement involves large extremities and often the entire body. Activities such as walking or sitting down involve gross motor capacities. Intentional movement, using fine and gross motor skills, involves a series of brain-based systems and is learned with repetition. With repeated action, the movement becomes rote or, as Luria (1973) described it, a “kinesthetic melody.”

Movement also can consist of both discrete and continuous patterns. Movements that are discrete might involve something as simple as lifting a finger, while continuous movements include an integrated set of skills like skipping. Movements may be disrupted if damage exists in the premotor cortex where the “kinesthetic melody” is believed to be formed. If this occurs, the individual may not be able to perform serial-continuous movements but may be able to demonstrate the specific discrete movements. Because of the complexity of motor patterns, the individual’s posture, movement in isolation, and movement in serial order should be assessed for possible intervention. This can be accomplished by observing individuals completing tasks such as writing their name (uses one hand), tying their shoes (uses both hands), and also performing novel tasks such as repeated tapping or clapping patterns. It should be noted if there is difficulty integrating the use of both hands. Both fine and gross motor skills should always be evaluated.

Although informal methods will yield a great deal of information regarding a client’s fine and gross motor functioning, several standardized instruments are also available which measure various specific or broad components of motoric functioning. For example, the *Bruininks–Oseretsky Test of Motor Proficiency* (Bruininks, 1978) provides a comprehensive picture of an individual’s motor development. The instrument was designed for children aged 4–5 to 14–15 and can be administered in 15–60 minutes, depending on whether the complete or short form of the battery is used. Three composite scores are provided in the areas of: Gross Motor Development (Running speed and agility, Balance, Bilateral coordination, Strength), Gross and Fine Motor Development (Upper-limb coordination), and Fine Motor Development (Response speed, Visual-motor control, Upper-limb speed and dexterity). Specific fine motor abilities can also be measured by using the Finger Tapping Test and the Grip Strength Test which are both a part of the Halstead-Reitan Battery (Reitan & Wolfson 1993).

To measure a client’s lateral preference, the *Lateral Preference Schedule* (Dean, 1988) can be administered to obtain a better understanding of clients’ lateral preference in the use of their eyes, ears, arms, hands, and feet (Rothlisberg, 1991). Atypical patterns of lateral preference have been hypothesized to indicate potential predictors of reading difficulty (Bemporad & Kinsbourne, 1983; Dean, Schwartz, & Smith, 1981). Determining lateral preference can be useful in interpreting assessment findings and in creating a plan for rehabilitation (Lezak, 1995).

4.09.4.3 Sensory-motor Integration

An additional component of our motoric functioning is the ability to integrate what is received by the senses with what is produced through action. For example, an individual may be able to perceive letters correctly and have adequate fine motor control, but still have difficulty correctly copying material presented in visual form. Numerous paper-and-pencil tests have been developed to assess motor function as it relates to visual-motor integration. Two of the most popular measures for this purpose are the *Bender Visual-Motor Gestalt* (Bender, 1938) and the *Developmental Test of Visual-Motor Integration (VMI)* (Beery, 1989).

The *Visual-Motor Gestalt Test* (Bender, 1938) is an individually administered test containing nine geometric figures which the client copies on to a blank sheet of paper. While historically this test was seen as a general measure of organicity, it is more appropriately used as a measure of visual-motor skills. Standard scores are provided in the developmental scoring system for children ages 5–0 to 11–11, although it is frequently used with adults as well. Most commonly known as the “Bender,” this measure is perhaps the best known and most widely used visual-motor assessment procedure available today (Bender, 1938; Reynolds & Kamphaus, 1990). As a component of a comprehensive assessment battery, performance on the Bender has long been thought to reveal visual-motor difficulties that may be associated with cerebral impairment (Sattler, 1992). Traditionally used to assess an individual’s constructional praxic skills, the Bender provides an evaluation of motor integration employed in the execution of complex learned movements (Hartlage & Golden, 1990). The information generated through this process may then be compared with levels of performance across other measures of functioning.

Alternate uses of the Bender include its administration as a memory test as well as a copying test. This dual administration process can be employed to assess different mental functions (short-term visual memory and visual perception) which utilize the same modalities in perception and task execution (Sattler, 1992). An additional technique available when interpreting the Bender performance is to have individuals compare the figure which they produced with the corresponding stimulus design. If the client is unable to recognize obvious differences between the two designs, a perceptual deficit may be involved. Likewise, if the client is able to detect a difference between the two figures, but is unable to make them alike, motor involvement may be influencing

performance (Hartlage & Golden, 1990). In the personality area, performance on the Bender may also be used to develop hypotheses regarding impaired performance due to poor planning, impulsivity, or compulsivity. Extremely large or small figures, heavily reinforced lines, and second attempts are examples of the item reproduction difficulties which are thought to indicate emotional concerns on the part of the individual.

The *VMI* is an individual or group administered test that involves copying a sequence of 24, increasingly complex, geometric figures. The test requires a relatively short administration time and is designed primarily for ages 4 to 13. The *VMI* offers several advantages as a tool for assessment and is widely used in psychological evaluation and research. The most common use of the *VMI*, now in its third edition, seems to be in assisting with the diagnosis of children who are suspected of having learning problems due to visual-motor difficulties. The *VMI* is also frequently employed when investigating the reliability and validity of other tests of visual-motor integration, such as the Bender, self drawing tasks, progressive matrices, and neuropsychological tests (Goldstein, Smith, & Waldrep, 1986; Palisano & Dichter, 1989). Because the *VMI* does not require a verbal response, it has also been used to assess visual-motor processes among non-English-speaking children (Brand, 1991; Frey & Pinelli, 1991).

4.09.4.3.1 Motor: implications for intervention

Motor problems and sensory-motor integration difficulties can impair a client’s ability to write or to learn new skills requiring motor coordination, and generally can have a negative impact on daily functioning. In the classroom setting, possible suggestions for accommodating these difficulties might include modifying instructions to compensate for motor difficulties (e.g., allowing pointing to the correct choice rather than writing, allowing students to tape record notes or copy them from others). In a rehabilitation setting, the client may need to learn alternative methods for writing such as using word processing programs on a computer or, if serious difficulties exist, using voice-activated programs. Consultation with an occupational therapist or a physical therapist will be extremely helpful in treatment planning when motor difficulties are evident.

4.09.4.4 Communication/Language

Language is the basic tool of human communication and hence essential to evaluate

when working with any client (Black & Strub, 1994). It should be viewed as a key skill because it serves as a primary means of conveying information from the individual to others and from others to the individual. Thus, communication difficulties have the power to influence all areas of life (D'Amato, Rothlisberg, & Rhodes, 1997). Difficulties or dysfunction found on tests of higher level functioning (e.g., learning processes) may well be secondary to a language disorder. Accordingly, language should be evaluated early in the course of an assessment to rule out problems in this area of functioning (Black & Strub, 1994; Lezak, 1995). Another obvious reason to evaluate language and communication skills is that language disorders occur as the result of a wide range of neurologic diseases and can manifest in a variety of forms of aphasia (e.g., Wernicke's, anomia, global, alexia, agraphia; Kolb & Whishaw, 1990). To aid in the interpretation of test findings, it is important for the clinician to be familiar with the various classic clinical aphasia presentations (see Gaddes & Edgell, 1994; Kolb & Whishaw, 1990; Lezak, 1995).

The language evaluation should be systematic and include an assessment of a range of relatively specific language functions. Assessment must evaluate both receptive and expressive verbal and nonverbal abilities to determine if adaptations are needed to enhance the individual in academic, vocational, and social situations. As part of an informal evaluation of expressive language, the clinician will want to evaluate spontaneous speech and verbal fluency by asking the client open-ended questions (Black & Strub, 1994). While the client is responding, the clinician can listen carefully for abnormal articulation, dysarthria (incoordination of the speech apparatus), verbal apraxia (difficulty carrying out purposeful speech), dysfluency, loss of prosody (melodic intonation), and disturbances of syntax or paraphasic errors (production of unintended syllables, words, or phrases) (Black & Strub, 1994; Kolb & Whishaw, 1990). Another important component of language, pragmatics, can also be evaluated. Pragmatics refers to the knowledge and activities of socially appropriate communication, which takes in much of the nonverbal aspects of communication, such as gestures, loudness of speech, as well as verbal appropriateness (Sohlberg & Mateer, 1990).

Evaluation of verbal fluency can be accomplished by counting the number of words the client is able to produce without repetition within a restricted category (e.g., animals or words beginning with a particular letter) and time (e.g., 60 seconds). The average adult should

produce approximately 20 animal names and a total of 40–60 words with performance depending to some degree on the client's intelligence, education, and social/linguistic background (Black & Strub, 1994). Additional methods of examining expressive language include having the client repeat back meaningful verbal phrases or sentences of increasing length and semantic complexity. Word finding and naming difficulties can be detected by the client's responses to the open-ended questions or by having the client describe a picture containing a series of objects or actions (Black & Strub, 1994).

Another major area of language functioning is receptive language or an individual's ability to understand what has been said. A comprehensive assessment should include an evaluation of the individual's ability to analyze and integrate information presented in a verbal format, since a common difficulty among those experiencing traumatic head injury is a decreased capacity to coordinate the social aspects of language (Ylvisaker, Szekeres, Haarbauer-Krupa, Urbanczyk, & Feeney, 1994). It is not sufficient to evaluate language comprehension based on open-ended questions because this method relies on expressive skills and does not examine comprehension in isolation (Black & Strub, 1994). Language comprehension can be evaluated informally by asking the client to point to common objects in the room or by asking a series of increasingly complex questions that require only a "yes" or "no" response (e.g., "Do dogs have four legs?") (Black & Strub, 1994). An evaluation of a client's reading and writing skills could also be included in an evaluation of language and communication (Black & Strub, 1994). The client can be asked to read sentences of increasing difficulty, spell words to dictation, and compose a paragraph in response to a prompt (e.g., "Tell me how to change a tire.") (Black & Strub, 1994).

In addition to the informal methods, several instruments are available that can prove useful for the clinical evaluation of language. There are a number of aphasia tests and batteries (e.g., *Boston Diagnostic Aphasia Examination*, Goodglass & Kaplan, 1983; *Multilingual Aphasia Examination*, Benton & Hamsher, 1989) which involve lengthy, well-controlled procedures and are best left to speech pathologists who have more extensive training in the specialized techniques of aphasia examinations (Lezak, 1995). As an alternative, aphasia screening tests can be used to indicate the presence of an aphasic disorder and may even highlight its specific characteristics, but do not provide the fine discriminations of the complete aphasia test batteries. Furthermore, these screening tests do not require technical knowledge of speech

pathology for adequate administration or determination of whether a significant aphasic disorder is present (Lezak, 1995). One of the most comprehensive aphasic screening tests available is the *Revised Token Test* (McNeil & Prescott, 1978). This expanded version of the original *Token Test* (De Renzi & Vignolo, 1962) contains 10 10-item subtests. McNeil and Prescott (1978) sought to ameliorate psychometric weaknesses of the original with this revision as well as seeking to develop an evaluative system for describing the nature and quantifying the degree of language deficit in order to facilitate treatment planning. Using tokens of various shapes and sizes, the clinician gives the client a series of increasingly complex instructions to follow. Though simple to administer, this instrument is reportedly very sensitive to disrupted linguistic processes that are central to the aphasic disability (Lezak, 1995).

Clinicians wanting a basic measure of different aspects of communication and language may wish to consider using the *Peabody Picture Vocabulary Test-Revised* (PPVT-R; Dunn & Dunn, 1981), the *Test of Language Development* (TOLD-2; Hammill & Newcomer, 1988), and the *Clinical Evaluation of Language Fundamentals-Revised* (Semel, Wiig, & Secord, 1987) or the *Test of Language Competence* (TLC; Wiig & Secord, 1989). Unfortunately, most of these tests are normed exclusively on children and adolescents and, therefore, have limited application to adults. One of the most frequently used tests, and one which has adult norms, is the *PPVT-R*. This test measures receptive vocabulary only and was normed for individuals aged two and a half to adulthood. The *PPVT-R* is untimed and requires the examinee to select from each plate of four pictures the one that best represents the target word. The test requires no reading ability, nor is the ability to point or provide an oral response essential (Shea, 1989). The *PPVT-R* can help to establish the level of verbal understanding a client has when expressive language is not required. Comparing such receptive skills with those expressive skills needed for other tests may help in developing hypotheses about the qualitative nature of verbal performance and in framing potential treatment (D'Amato, Rothlisberg, & Rhodes, 1997).

The *TOLD-2* is available in a form designed for primary ages (4–0 to 8–11) and intermediate ages (8–6 to 12–11). This test purports to measure receptive and expressive language proficiency. The results for the *TOLD-2* provide quotients for an Overall Spoken Language score, and for the composites of Listening, Speaking, Semantics, Syntax, and Phonology

(on the primary form only). Overall, the *TOLD-2* instruments are reliable and valid as language screening tools for younger clients (Wochnick Fodness, McNeilly, & Bradley-Johnson, 1991; Westby, 1988). Some tests have been developed to measure more complex language usage, such as the *TLC* (Wiig & Secord, 1989) which purports to measure metalinguistic abilities. The four subtests involve producing multiple meanings for ambiguous sentences, recognizing inferences on the basis of incomplete information, creating sentences given three words and a context, and recognizing the meaning of figurative language. This type of test may be useful for identifying subtle problems in language usage (Crosson, 1996).

4.09.4.4.1 Communication/language: implications for intervention

For clients who are experiencing difficulty with either or both receptive and expressive language, the clinician can modify verbal interaction by shortening the length of information presented or presenting information in steps. Additional ideas for the school-age client might include recommending that the teacher repeat directions and have the student also repeat and explain the directions back to the teacher, pairing verbal instructions with non-verbal cues, and using nonverbal cues. Many of these strategies could be adapted to adults in rehabilitation and other types of therapeutic settings as well. The clinician must be careful to check frequently with clients to ensure understanding and to assist the clients and their families to adjust to these communication or language deficits.

4.09.5 FUTURE DIRECTIONS

Although knowledge about how individuals process information has grown exponentially since the late 1970s, researchers and practitioners alike are left with many questions regarding how individuals remember and learn. How do age, gender, and ethnicity impact a client's functioning on these specific instruments? Do individuals with brain damage process information differently than individuals with "normal" brain functioning? How do the results of an assessment translate into effective treatment strategies that will help individuals function better on the job? Despite these questions and more, as a field we do know that current measures of processing can allow practitioners to make predictions with a reasonable level of confidence. However, we must also recognize that future research investigating the

prediction accuracy of various tests is needed to expand the range and precision of clinical prediction (Long, 1996). Indeed, limited research exists that evaluates the effect of various treatment approaches on success in clinical pediatric or adult populations (Batchelor, 1996b; Ris & Noll, 1994). As our knowledge of the nervous system is expanded, and we begin to understand the intricacies of the brain's organization, we can begin to see how perceptions are formed, information stored and integrated, and action taken. Until that time, the explanation for behavior and certain learning difficulties can only be inferred (D'Amato, Rothlisberg, & Leu, *in press*).

Another component complicating our enhanced understanding of information processing is the need for common language and goals between neurologists, psychologists, educational researchers, and vocational rehabilitation specialists. Bigler (1996) notes that it is

essential that physicians and psychologists work toward some common understanding of normal and abnormal behavioral manifestations of brain functioning, particularly aspects of complex attention and integration of sensory experiences, memory, motivation, organization of verbal and nonverbal cognition, abstract thinking, problem solving, executive functions, and self-monitoring of behavior. Without agreement on a detailed and relatively comprehensive model of neurobehavioral development, psychologists will be limited in their ability to develop appropriate assessment and intervention strategies. (p. 50)

Unfortunately, outcomes in rehabilitation research have also suffered from a lack of consensus with importance placed on different variables depending on the orientation of the author and audience (Batchelor, 1996b). For example, clinical researchers have examined pre- and postperformance measures of cognitive, motivational, and behavioral functions (Ris & Noll, 1994), while service providers have focused on outcome constructs such as employment and independent living (e.g., Adunsky, Hershkowitz, Rabbi, Asher-Sivron, & Ohry, 1992). Concurrently, third-party payers are interested in length of stay, cost, and effectiveness in allocation of resources (Fratalli, 1993), while families and consumers are interested in quality of care. Batchelor (1996b) concluded that the majority of research has emphasized short-term outcomes and the meaningful questions generated by service providers, consumers, and third-party payers have been overlooked and present an important direction for the field to pursue.

In terms of the practical aspects of assessment, the practitioner's task is not an easy one.

In order to generate an accurate diagnosis or provide the most sound recommendations, a complete understanding of the client is necessary. A practitioner could spend hours evaluating each of the areas outlined with careful consideration of all subdomains using a multidimensional approach. Although this approach would yield a bounty of information, it may not be practical given time limitations and insurance policy guidelines. The key for the clinician is to find a balance between finding out the most important information about client functioning through the use of instruments with the best predictive ability and spending a limited amount of time on assessment. By creating an efficient and effective assessment approach, more time is available to implement treatment. The task of generating recommendations for interventions that are likely to enhance client functioning is of central importance to the issue of assessment. To this end, continued information is needed on the relationship between assessed cognitive processing and predicted future performance in real-world settings (Sbordone & Long, 1996). Furthermore, we must gain knowledge about the most effective intervention strategies for all types of individuals. Future research will enable us to understand how the brain processes information and what treatments are effective with what types of clients. Indeed, this additional knowledge may allow us to match client subtypes with specific treatments which will increase the effectiveness and efficacy of psychological services.

4.09.6 SUMMARY

The assessment of children, adolescents, and adults encompasses a wide range of domains from which a clinician may view a client's functioning. Consideration must be given to the many layers of the client context (e.g., family support, socioeconomic status, domain of functioning), client characteristics (e.g., motivation, education level, ethnicity, age), as well as the specific cognitive processes under question (e.g., memory, sensory perception). Although often overlooked or subsumed within the broader arenas of intelligence or achievement, the specific areas of memory, learning, and special aptitudes are critical to our daily functioning. That is, one will have difficulty demonstrating intelligence or learning new tasks, if there is a severe memory deficit or difficulty in accurately perceiving stimuli. The descriptions of the domains presented in this chapter have offered insight into the field's current understanding of these systems as well as the breadth of evaluation strategies available

to probe the diverse nature of cognitive processes. Once the practitioner identifies the assessment needs of the individual client, and a decision is made as to the components most relevant for exploring the referral question, the process can begin. By generating quality data, our ability to predict outcomes and provide effective potential intervention strategies is increased. Indeed, the goal of any assessment is to respond correctly to the question presented by the referral source, provide accurate predictions of future outcomes, and generate effective strategies for improving the client's adaptation or functioning.

4.09.7 REFERENCES

- Adunsky, A., Hershkowitz, M., Rabbi, R., Asher-Sivron, L., & Ohry, A. (1992). Functional recovery in young stroke patients. *Archives of Physical Medicine and Rehabilitation*, 73, 859–862.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington DC: Author.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Arter, J. A., & Jenkins, J. R. (1977). Examining the benefits and prevalence of modality considerations in special education. *The Journal of Special Education*, 11, 291–298.
- Baddeley, A. (1986). *Working memory*. Oxford, UK: Oxford University Press.
- Baddeley, A., Emslie, H., & Nimmo-Smith, I. (1994). *Doors and People: A Test of Visual and Verbal Recall and Recognition*. Suffolk, UK: Thames Valley Test Co.
- Baldwin, J. D., & Baldwin, J. I. (1986). *Behavior principles in everyday life* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Barkley, R. A. (1996). Critical issues in research on attention. In G. R. Lyon & N. A. Krasnegor (Eds.), *Attention, memory, and executive function* (pp. 45–56). Baltimore: Brookes.
- Batchelor, E. S. (1996a). Neuropsychological assessment of children. In E. S. Batchelor & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation* (pp. 9–26). Boston: Allyn & Bacon.
- Batchelor, E. S. (1996b). Future considerations for rehabilitation research and outcome studies. In E. S. Batchelor & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation* (pp. 347–352). Boston: Allyn & Bacon.
- Beery, K. E. (1989). *Developmental Test of Visual-Motor Integration*. Odessa, FL: Psychological Assessment Resources.
- Begali, V. (1992). *Head injury in children and adolescents: A resource and review for school and allied professionals* (2nd ed.). Brandon, VT: Clinical Psychology Publishing Company.
- Begali, V. (1994). The role of the school psychologist. In R. C. Savage & G. F. Wolcott (Eds.), *Educational dimensions of acquired brain injury* (pp. 453–473). Austin, TX: PRO-ED.
- Bemporad, B., & Kinsbourne, M. (1983). Sinistrality and dyslexia: A possible relationship between subtypes. *Topics in Learning and Learning Disabilities*, 3(1), 48–65.
- Bender, L. (1938). A visual motor gestalt test and its clinical use. *American Orthopsychiatric Association Research Monograph*, No. 3. New York: American Orthopsychiatric Association.
- Benton, A. L., & Hamsher, K. deS. (1989). *Multilingual Aphasia Examination*. Iowa City, IA: AJA Associates.
- Benton, A. L., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment*. New York: Oxford University Press.
- Benton-Sivan, A. (1992). *The Revised Visual Retention Test* (5th ed.). New York: The Psychological Corporation.
- Berk, R. A. (1995). Review of the Memory Assessment Scale. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurement yearbook* (pp. 593–594). Lincoln, NE: Buros.
- Bigler, E. D. (1996). Bridging the gap between psychology and neurology: Future trends in pediatric neuropsychology. In E. S. Batchelor & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation* (pp. 27–54). Boston: Allyn & Bacon.
- Black, F. W., & Strub, R. L. (1994). The bedside and office mental status examination. In S. Touyz, D. Byrne, & A. Gilandas (Eds.), *Neuropsychology in clinical practice* (pp. 38–60). Boston: Academic Press.
- Brand, H. J. (1991). Correlation for scores on revised tests of visual-motor integration and copying test in a South African sample. *Perceptual and Motor Skills*, 73, 225–226.
- Brown, R. T., Armstrong, F. D., & Eckman, J. R. (1993). Neurocognitive aspects of pediatric sickle cell disease. *Journal of Learning Disabilities*, 26, 33–45.
- Brown, V. L., Hammill, D. D., & Wiederholt, J. L. (1995). *Test of Reading Comprehension-3 (TORC-3)*. Austin, TX: PRO-ED.
- Bruininks, R. H. (1978). *Bruininks-Oseretsky Test of Motor Proficiency*. Circle Pines, MN: American Guidance Service.
- Campbell, J. W., D'Amato, R. C., Raggio, D. J., & Stephens, K. D. (1991). Construct validity of the computerized Continuous Performance Test with measures of intelligence, achievement, and behavior. *Journal of School Psychology*, 29, 143–150.
- Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: Foundations and fundamentals* (pp. 82–115). New York: Guilford.
- Cantor, J., Engle, R. W., & Hamilton, G. (1991). Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence*, 15, 229–246.
- Cohen, S. B. (1991). Adapting educational programs for students with head injuries. *Journal of Head Trauma Rehabilitation*, 1, 56–63.
- Colarusso, R. P., & Hammill, D. D. (1996). *Motor-Free Visual Perception Test-Revised (MFPT-R)*. Novato, CA: Academic Therapy.
- Cole E., & Siegel, J. A. (1990). School psychology in a multicultural community: Responding to children's needs. In E. Cole & J. A. Siegel (Eds.), *Effective consultation in school psychology* (pp. 141–169). Toronto, ON: Hogrefe & Huber.
- Colvin, S. S. (1921). Intelligence and its measurement: A symposium (IV). *Journal of Educational Psychology*, 12, 136–139.
- Connolly, A. J. (1988). *Keymath-revised: A diagnostic inventory of essential mathematics*. Circle Pines, MN: American Guidance Service.
- Cooley, E. L., & Morris, R. D. (1990). Attention in children: A neuropsychology based model of assessment. *Developmental Neuropsychology*, 6, 239–274.
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods. A handbook for research on interactions*. New York: Irvington.
- Crosson, B. (1996). Assessment of subtle language deficits in neuropsychological batteries: Strategies and implications. In R. J. Sbordone & C. J. Long (Eds.), *Ecological*

- validity of neuropsychological testing (pp. 243–259). Delray Beach, FL: GR Press/St Lucie Press.
- D'Amato, R. C. (1990). A neuropsychological approach to school psychology. *School Psychology Quarterly*, 5, 141–160.
- D'Amato, R. C., & Dean, R. S. (Eds.) (1989a). *The school psychologist in nontraditional settings: Integrating clients, services, and settings*. Hillsdale, NJ: Erlbaum.
- D'Amato, R. C., & Dean, R. S. (1989b). The past, present, and future of school psychology in nontraditional settings. In R. C. D'Amato & R. S. Dean (Eds.), *The school psychologist in nontraditional settings: Integrating clients, services, and settings* (pp. 185–209). Hillsdale, NJ: Erlbaum.
- D'Amato, R. C., & Rothlisberg, B. A. (1992). *Psychological perspectives on intervention: A case study approach to prescriptions for change*. New York: Longman.
- D'Amato, R. C., & Rothlisberg, B. A. (1996). How education should respond to students with traumatic brain injuries. *Journal of Learning Disabilities*, 29, 670–683.
- D'Amato, R. C., Rothlisberg, B. A., & Leu, P. W. (in press). Neuropsychological assessment for intervention. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed.). New York: Wiley.
- D'Amato, R. C., Rothlisberg, B. A., & Rhodes, R. L. (1997). Utilizing a neuropsychological paradigm for understanding common educational and psychological tests. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed.). New York: Plenum.
- Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn & Bacon.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Das, J. P., Kirby, J., & Jarman, R. F. (1979). *Simultaneous and successive cognitive processes*. New York: Academic Press.
- Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes. The PASS theory of intelligence*. New York: Allyn & Bacon.
- Dean, R. S. (1977). Canonical analysis of a jangle fallacy. *Multivariate Experimental Clinical Research*, 3, 17–20.
- Dean, R. S. (1983). Intelligence-achievement discrepancies in diagnosing pediatric learning disabilities. *Clinical Neuropsychology*, 3, 58–62.
- Dean, R. S. (1984). Functional lateralization of the brain. *Journal of Special Education*, 18, 239–256.
- Dean, R. S. (1985a). Neuropsychological assessment. In R. Michels, J. O. Cavenar, H. K. H. Brodie, A. M. Cooper, S. B. Guze, L. L. Judd, G. L. Klerman, & A. J. Solnit (Eds.), *Psychiatry* (pp. 1–16). Philadelphia: Lippincott.
- Dean, R. S. (1985b). Foundation and rationale for neuropsychological bases of individual differences. In L. C. Hartlage & C. F. Telzrow (Eds.), *The neuropsychology of individual differences: A developmental perspective* (pp. 7–39). New York: Plenum.
- Dean, R. S. (1986). Perspectives on the future of neuropsychological assessment. In B. S. Plake & J. C. Witt (Eds.), *Buros-Nebraska series on measurement and testing: Future of testing and measurement* (pp. 203–241). Hillsdale, NJ: Erlbaum.
- Dean, R. S. (1988). *Lateral Preference Schedule*. Odessa, FL: Psychological Assessment Resources.
- Dean, R. S., & Gray, J. W. (1990). Traditional approaches to neuropsychological assessment. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 371–388). New York: Guilford Press.
- Dean, R. S., Schwartz, N. H., & Smith, L. S. (1981). Lateral preference patterns as a discriminator of learning difficulties. *Journal of Consulting and Clinical Psychology*, 49, 227–235.
- Dearborn, W. F. (1921). Intelligence and its measurement: A symposium (XII). *Journal of Educational Psychology*, 12, 210–212.
- De Renzi, E., & Vignolo, L. A. (1962). The Token Test: A sensitive test to detect disturbances in aphasics. *Brain*, 85, 665–678.
- Drew, R. H., & Templer, D. I. (1992). Contact sports. In D. I. Templer, L. C. Hartlage, & W. G. Cannon (Eds.), *Preventable brain damage: Brain vulnerability and health* (pp. 15–29). New York: Springer.
- Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised*. Circle Pines, MN: American Guidance Service.
- Eliason, M. J., & Richman, L. C. (1987). The Continuous Performance Test in learning disabled and nondisabled children. *Journal of Learning Disabilities*, 20, 614–619.
- Feuerstein, R., Rand, Y., & Hoffman, M. (1979). *The dynamic assessment of retarded performers: The Learning Potential Assessment Device: Theory, instruments, and techniques*. Baltimore: University Park.
- Figueroa, R. A., & Garcia, E. (1994). Issues in testing students from culturally and linguistically diverse backgrounds. *Multicultural Education*, 2, 10–19.
- Frattali, C. M. (1993). Perspectives on functional assessment: Its use for policy making. *Disability and Rehabilitation*, 15, 1–9.
- Frey, P. D., & Pinelli, B. (1991). Visual discrimination and visuomotor integration among two classes of Brazilian children. *Perceptual and Motor Skills*, 72, 847–850.
- Fuchs, L. S. (1994). Integrating curriculum-based measurement with instructional planning for students with learning disabilities. In N. C. Jordan & J. Goldsmith-Phillips (Eds.), *Learning disabilities: New directions for assessment and intervention* (pp. 177–195). Boston: Allyn & Bacon.
- Gaddes, W. H., & Edgell, D. (1994). *Learning disabilities and brain function: A neuropsychological approach* (3rd ed.). New York: Springer-Verlag.
- Geil, M., & D'Amato, R. C. (1996). *Contemporary ecological neuropsychology: An alternative to the medical model for conceptualizing learning disabilities*. Manuscript submitted for publication.
- Golden, C. J. (1981). The Luria-Nebraska Children's Battery: Theory and formulation. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-aged child: Issues and procedures* (pp. 277–302). New York: Grune & Stratton.
- Golden, C. J., Sawicki, R. F., & Franzen, M. D. (1984). Test construction. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 19–37). New York: Pergamon.
- Goldstein, D. J., Smith, K. B., & Waldrep, E. E. (1986). Factor analytic study of the Kaufman Assessment Battery for Children. *Journal of Clinical Psychology*, 42, 890–894.
- Goodglass, H., & Kaplan, E. (1983). *Boston Diagnostic Aphasia Examination (BDAAE)*. Philadelphia: Lea and Febiger. Distributed by Psychological Assessment Resources, Odessa, FL.
- Gray, J. W., & Dean, R. S. (1989). Approaches to the cognitive rehabilitation of children with neuropsychological impairment. In C. R. Reynolds & F. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 397–408). New York: Plenum.
- Greenberg, L. (1993). *Test of variables of attention (T.O.V.A.™)*. Wood Dale, IL: Stoetling.
- Greenberg, L. M., & Waldman, I. D. (1993). Developmental normative data on the test of variables of attention (T.O.V.A.™). *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 34, 1019–1030.
- Guilmette, T. J., & Giuliano, A. J. (1991). Taking the

- stand: Issues and strategies in forensic neuropsychology. *The Clinical Neuropsychologist*, 5, 197–219.
- Gutkin, T. B., & Reynolds, C. R. (Eds.) (1990). *The handbook of school psychology* (2nd ed.). New York: Wiley.
- Halperin, J. M., Sharma, V., Greenblatt, E., & Schwartz, S. (1991). Assessment of the Continuous Performance Test: Reliability and validity in a nonreferred sample. *Psychological Assessment*, 3, 603–608.
- Hammill, D. D. (1991). *Detroit Tests of Learning Aptitude (DTLA-3)* (3rd ed.). Austin, TX: PRO-ED.
- Hammill, D. D., & Bryant, B. R. (1991a). *Detroit Tests of Learning Aptitude-Adult (DTLA-A)*. Austin, TX: PRO-ED.
- Hammill, D. D., & Bryant, B. R. (1991b). *Detroit Tests of Learning Aptitude-Primary (DTLA-P:2)* (2nd ed.). Austin, TX: PRO-ED.
- Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language-3 (TOWL-3)*. Austin, TX: PRO-ED.
- Hammill, D. D., & Newcomer, P. L. (1988). *Test of Language Development Intermediate (TOLD-2)* (2nd ed.). Austin, TX: PRO-ED.
- Hamsher, K. de S. (1984). Specialized neuropsychological assessment methods. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 235–256). New York: Pergamon.
- Hartlage, L. C., & Golden, C. J. (1990). Neuropsychological assessment techniques. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed., pp. 431–457). New York: Wiley.
- Hartlage, L. C., & Telzow, C. F. (1983). The neuropsychological basis of educational intervention. *Journal of Learning Disabilities*, 16, 521–528.
- Hooper, S. R. (1995). Review of the Visual Search and Attention Test. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook* (pp. 1081–1082). Lincoln, NE: Buros.
- Huebner, E. S. (1992). Review of the Wechsler Memory Scale-Revised. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurement yearbook* (pp. 1023–1024). Lincoln, NE: Buros.
- Hynd, G. W., & Semrud-Clikeman, M. (1990). Neuropsychological assessment. In A. S. Kaufman (Ed.), *Assessing adolescent and adult intelligence* (pp. 638–695). Boston: Allyn & Bacon.
- Hynd, G. W., & Willis, W. G. (1988). *Pediatric neuropsychology*. Boston: Allyn & Bacon.
- Jarvis, P. E., & Barth, J. T. (1994). *The Halstead-Reitan Neuropsychological Battery: A guide to interpretation and clinical applications*. Odessa, FL: Psychological Assessment Resources.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122–149.
- Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence*. Boston: Allyn & Bacon.
- Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: Wiley.
- Kavale, K. A., Forness, R. F., & Bender, M. (1988). *Handbook of learning disabilities: Volume II: Methods and interventions*. Boston: College-Hill.
- Klee, S. H., & Garfinkel, B. D. (1983). The computerized Continuous Performance Task: A new measure of inattention. *Journal of Abnormal Child Psychology*, 11, 489–495.
- Kolb, B., & Whishaw, I. Q. (1990). *Fundamentals of human neuropsychology* (3rd ed.). New York: Freeman.
- Kovacs, M., Goldston, D., & Ivengar, S. (1992). Intellectual development and academic performance of children with insulin-dependent diabetes mellitus: A longitudinal study. *Developmental Psychology*, 28, 676–684.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14, 389–433.
- LaBerge, D. (1995). *Attentional processing: The brain's art of mindfulness*. Cambridge, MA: Harvard University Press.
- Lassiter, K. S., D'Amato, R. C., Raggio, D. J., Whitten, J. C. M., & Bardos, A. N. (1994). The construct specificity of the Continuous Performance Test: Does inattention relate to behavior and achievement? *Developmental Neuropsychology*, 10, 179–188.
- Lezak, M. D. (1983). *Neuropsychological assessment* (2nd ed.). New York: Oxford University Press.
- Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.
- Lindgren, S. D., & Lyon, D. (1983). *PACE: Pediatric assessment of cognitive efficiency*. Iowa City, IA: University of Iowa, Department of Pediatrics.
- Long, C. J. (1996). Neuropsychological tests: A look at our past and the impact that ecological issues may have on our future. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 1–14). Delray Beach, FL: GR Press/St Lucie Press.
- Luria, A. R. (1970). The functional organization of the brain. *Scientific American*, 222(3), 66–78.
- Luria, A. R. (1973). *The working brain: An introduction to neuropsychology*. New York: Basic Books.
- Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.
- Markwardt, (1989). *Peabody Individual Achievement Test-Revised (PIAT-R)*. Circle Pines, MN: American Guidance Service.
- Martinez, M. A. (1985). Toward a bilingual school psychology model. *Educational Psychology*, 20, 143–152.
- Mastropieri, M. A., & Scruggs, T. E. (1989). Constructing more meaningful relationships: Mnemonic instruction for special populations. *Educational Psychology Review*, 1, 83–111.
- McNeil, M. M., & Prescott, T. E. (1978). *Revised Token Test*. Austin, TX: PRO-ED.
- Morris, R. D. (1996). Relationships and distinctions among the concepts of attention, memory, and executive function: A developmental perspective. In G. R. Lyon & N. A. Krasnegor (Eds.), *Attention, memory, and executive function* (pp. 11–16). Baltimore: Brookes.
- Palincsar, A., Brown, A. L., & Campione, J. C. (1991). Dynamic assessment. In H. L. Swanson (Ed.), *Handbook on the assessment of learning disabilities: Theory, research, and practice* (pp. 75–95). Austin, TX: PRO-ED.
- Palisano, R. J., & Dichter, C. G. (1989). Comparison of two tests of visual-motor development used to assess children with learning disabilities. *Perceptual and Motor Skills*, 68, 1099–1103.
- Poteat, G. M. (1995). Review of the Detroit Tests of Learning Aptitude, Third Edition. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurement yearbook* (pp. 277–278). Lincoln, NE: Buros.
- Pressley, M., & Levin, J. R. (Eds.) (1983). *Cognitive strategy research: Psychological foundations*. New York: Springer-Verlag.
- Raggio, D. (1991). *Raggio Evaluation of Attention Deficit Disorder (Computerized test)*. Jackson, MS: University of Mississippi Medical Center, Infant and Child Development Clinic.
- Reitan, R. M., & Davison, L. A. (1974). *Clinical neuropsychology: Current status and applications*. New York: Winston/Wiley.
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

- Reitan, R. M., & Wolfson, D. (1995, October). *Cognitive and emotional consequences of mild head injury*. Paper presented at the fall conference of the Colorado Neuropsychological Society, Colorado Springs, CO.
- Resnick, L. B. (Ed.) (1976). *The nature of intelligence*. Hillsdale, NJ: Erlbaum.
- Reynolds, C. R. (1981a). The neuropsychological basis of intelligence. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-aged child: Issues and procedures* (pp. 87–124). New York: Grune & Stratton.
- Reynolds, C. R. (1981b). Neuropsychological assessment and the habilitation of learning: Considerations in the search for the aptitude x treatment interaction. *School Psychology Review*, 10, 343–349.
- Reynolds, C. R. (1986). Transactional models of intellectual development, yes. Deficit models of process remediation, no. *School Psychology Review*, 15, 256–260.
- Reynolds, C. R., & Bigler, E. D. (1994). *Test of memory and learning*. Austin, TX: PRO-ED.
- Reynolds, C. R., & Kamphaus, R. W. (1990). *Handbook of psychological and educational assessment of children: Intelligence and achievement*. New York: Guilford Press.
- Ris, D., & Noll, R. B. (1994). Long-term neurobehavioral outcome in pediatric brain-tumor patients: Review and methodological critique. *Journal of Clinical and Experimental Neuropsychology*, 16(1), 21–42.
- Rosvold, H., Mirsky, A., Sarason, I., Bransome, L., & Beck, L. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, 20, 343–350.
- Rothlisberg, B. A. (1991). Factor stability of the Lateral Preference Schedule. *International Journal of Neuroscience*, 61, 83–85.
- Rothlisberg, B. A. (1992). Integrating psychological approaches to intervention. In R. C. D'Amato & B. A. Rothlisberg (Eds.), *Psychological perspectives on intervention: A case study approach to prescriptions for change* (pp. 190–198). New York: Longman.
- Rovet, J. F., Ehrlich, R. M., Czuchta, D., & Akler, M. (1993). Psychoeducational characteristics of children and adolescents with insulin-dependent diabetes mellitus. *Journal of Learning Disabilities*, 26, 7–22.
- Rovet, J. F., Ehrlich, R. M., & Hoppe, M. (1988). Specific intellectual deficits in children with early onset diabetes mellitus. *Child Development*, 59, 226–234.
- Sattler, J. M. (1992). *Assessment of children* (3rd ed., rev.). San Diego, CA: Sattler.
- Sbordone, R. J., & Long, C. J. (Eds.) (1996). *Ecological validity of neuropsychological testing*. Delray Beach, FL: GR Press/St Lucie Press.
- Selz, M. (1981). Halstead-Reitan neuropsychological test batteries for children. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-aged child: Issues and procedures* (pp. 195–235). New York: Grune & Stratton.
- Semel, E., Wiig, E. H., & Secord, W. (1987). *Clinical Evaluation of Language Fundamentals-Revised (CELF-R)*. San Antonio, TX: Psychological Corp.
- Shea, V. (1989). Peabody Picture Vocabulary Test-Revised. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (Vol. II, pp. 271–283). Boston: Allyn & Bacon.
- Sheslow, D., & Adams, W. (1990). *Wide Range Assessment of Memory and Learning (WRAML)*. Wilmington, DE: Jastak.
- Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review*, 76, 179–193.
- Shinn, M. R. (Ed.) (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.
- Silver, L. B. (1993). Introduction and overview to the clinical concepts of learning disabilities. *Child and Adolescent Psychiatric Clinics of North America: Learning Disabilities*, 2, 181–192.
- Simpson, N., Black, F. W., & Strub, R. L. (1986). Memory assessment using the Strub-Black mental status examination and the Wechsler Memory Scale. *Journal of Clinical Psychology*, 42, 147–155.
- Slomka, G. T., & Tarter, R. E. (1993). Neuropsychological assessment. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child and adolescent assessment* (pp. 208–223). Boston: Allyn and Bacon.
- Sohlberg, M. M., & Mateer, C. A. (1990). Evaluation and treatment of communicative skills. In J. S. Kreutzer & P. Wehman (Eds.), *Community integration following traumatic brain injury*. Baltimore: Paul H. Brookes.
- Strub, R. L., & Black, F. W. (1993). *The mental status examination in neurology* (3rd ed.). Philadelphia: F. A. Davis.
- Swanson, H. L. (1981). Vigilance deficits in learning disabled children: A signal detection analysis. *Journal of Psychology and Psychiatry*, 2, 339–398.
- Swanson, H. L. (1995). Using the Cognitive Processing Test to assess ability: Development of a dynamic assessment measure. *School Psychology Review*, 24, 672–693.
- Swanson, H. L. (1996). *Swanson Cognitive Processing Test (S-CPT)*. Austin, TX: PRO-ED.
- Talley, J. L. (1993). *Children's Auditory Verbal Learning Test-2 (CAVLT-2)*. Odessa, FL: Psychological Assessment Resources.
- Tarver, S. G., & Dawson, M. M. (1978). Modality preference and the teaching of reading: A review. *Journal of Learning Disabilities*, 11, 5–17.
- Taylor, H. G. (1988). Learning disabilities. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders* (2nd ed., pp. 402–450). New York: Guilford Press.
- Taylor, H. G., & Fletcher, J. M. (1990). Neuropsychological assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed., pp. 228–255). New York: Pergamon.
- Taylor, H. G., Fletcher, J. M., & Satz, P. (1984). Neuropsychological assessment in children. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 211–234). New York: Pergamon.
- Telzrow, C. F. (1985). The science and speculation of rehabilitation in developmental neuropsychological disorders. In L. C. Hartlage & C. F. Telzrow (Eds.), *The neuropsychology of individual differences: A developmental perspective* (pp. 271–307). New York: Plenum.
- Templer, D. I., & Drew, R. H. (1992). Noncontact sports. In D. I. Templer, L. C. Hartlage, & W. G. Cannon (Eds.), *Preventable brain damage: Brain vulnerability and health* (pp. 30–40). New York: Springer.
- Touyz, S., Byrne, D., & Gilandas, A. (1994). *Neuropsychology in clinical practice*. Boston: Academic Press.
- Trenerry, M. R., Crosson, B., DeBoe, J., & Leber, W. R. (1990). *Visual search and attention test*. Odessa, FL: Psychological Assessment Resources.
- Walsh, K. W. (1978). *Neuropsychology: A clinical approach*. New York: Churchill Livingstone.
- Wechsler, D. (1987). *Wechsler Memory Scale-Revised manual*. San Antonio, TX: The Psychological Corporation.
- Wedding, D., Horton, A. M., & Webster, J. S. (1986). *The neuropsychology handbook: Behavioral and clinical perspectives*. New York: Springer.
- Wepman, J. M., & Reynolds, W. M. (1987). *Wepman's Auditory Discrimination Test* (2nd ed.) Los Angeles: Western Psychological Services.
- Westby, C. (1988). Test review: Test of Language Development-2 Primary, Test of Language Development-2 Intermediate. *The Reading Teacher*, 42, 236–237.
- Whitten, J. C., D'Amato, R. C., & Chittooran, M. M.

- (1992). A neuropsychological approach to intervention. In R. C. D'Amato & B. A. Rothlisberg (Eds.), *Psychological perspectives on intervention: A case study approach to prescriptions for change* (pp. 112–136). White Plains, NY: Longman.
- Wiig, E. H., & Second, W. (1989). *Test of Language Competence-Expanded Edition (TLC)*. San Antonio, TX: The Psychological Corporation.
- Williams, J. M. (1991). *Memory Assessment Scales (MAS)*. Odessa, FL: Psychological Assessment Resources.
- Williams, R. E., & Vincent, K. R. (1991). Review of the Peabody Individual Achievement Test-Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques* (Vol. 8, pp. 557–562). Kansas City, MO: Test Corporation of America.
- Wochnick Fodness, R., McNeilly, J., & Bradley-Johnson, S. (1991). Test-retest reliability of the Test of Language Development-2: Primary and Test of Language Development-2: Intermediate. *Journal of School Psychology*, 29, 161–165.
- Woodcock, R., & Johnson, M. B. (1989). *Woodcock-Johnson Psychoeducational Battery-Revised (WJPB-R)*. Chicago: Riverside.
- Woodrow, H. (1921). Intelligence and its measurement: A symposium (XI). *Journal of Educational Psychology*, 12, 207–210.
- Ylvisaker, M., Szekeres, S. F., Haarbauer-Krupa, J., Urbanczyk, B., & Feeney, T. J. (1994). Speech and language intervention. In R. C. Savage & G. F. Wolcott (Eds.), *Educational dimensions of acquired brain injury* (pp. 185–235). Austin, TX: PRO-ED.
- Ylvisaker, M., Szekeres, S. F., & Hartwick, P. (1994). A framework for cognitive intervention. In R. C. Savage & G. F. Wolcott (Eds.), *Educational dimensions of acquired brain injury* (pp. 35–67). Austin, TX: PRO-ED.
- Youngjohn, J. R., Larrabee, G. J., & Crook, T. H. (1993). New adult age- and education-correction norms for the Benton Visual Retention Test. *The Clinical Neuropsychologist*, 7, 155–160.

4.10

Neuropsychological Assessment of Children

CYNTHIA A. RICCIO and CECIL R. REYNOLDS
Texas A&M University, College Station, TX, USA

4.10.1 INTRODUCTION	267
4.10.1.1 <i>Assessment Process</i>	269
4.10.2 MEASURES USED IN THE ASSESSMENT OF CHILDREN	271
4.10.2.1 <i>Neuropsychological Interpretation of Children's Measures</i>	272
4.10.2.2 <i>Development of New Measures for Children</i>	273
4.10.2.3 <i>Current and Future Trends</i>	274
4.10.2.3.1 <i>Memory</i>	274
4.10.2.3.2 <i>Attention</i>	275
4.10.2.3.3 <i>Computer-administered assessment</i>	276
4.10.2.3.4 <i>Integration of neuroimaging and electrophysiology</i>	276
4.10.2.3.5 <i>Integration of cognitive and developmental psychology</i>	277
4.10.2.4 <i>Measurement Issues</i>	277
4.10.2.5 <i>Approaches to Test Selection with Children</i>	279
4.10.2.5.1 <i>Nomothetic approaches</i>	280
4.10.2.5.2 <i>Idiographic approaches</i>	283
4.10.2.5.3 <i>Combined approaches</i>	285
4.10.2.6 <i>General Organization of the Neuropsychological Assessment of the Child</i>	287
4.10.2.7 <i>Interpretation Issues</i>	289
4.10.2.7.1 <i>Performance level</i>	290
4.10.2.7.2 <i>Profile patterns</i>	290
4.10.2.7.3 <i>Functional asymmetry</i>	290
4.10.2.7.4 <i>Pathognomonic signs</i>	291
4.10.2.7.5 <i>Combination approaches</i>	291
4.10.3 CONCLUSIONS	291
4.10.4 REFERENCES	293

4.10.1 INTRODUCTION

The area of clinical neuropsychology has only recently been established as a viable specialty area (Woody, 1997). By definition, neuropsychology is the study of brain-behavior relationships that uses the theory and methodologies of both neurology and psychology. Historically, neuropsychology has been used for the diagnostic assessment of adults with known brain damage or injury; the clinical research has been

derived from the study of adults with identified insult to the brain. The major premise of neuropsychological assessment is that different behaviors, including higher order cognitive skills, involve differing neurological structures or functional systems (Luria, 1980). As such, the neuropsychological approach to assessment involves assessment of various behavioral domains believed to be related to functional systems and making inferences about brain integrity based on the individual's performance

across these domains. Neuropsychological assessment samples behaviors known to depend on the integrity of the central nervous system (CNS) using measures that correlate with cognitive, sensorimotor, and emotional functioning based on clinical research (Dean & Gray, 1990).

As more became known about brain-behavior relationships, clinical findings and theories were applied to the understanding of learning and behavior problems of adults where brain damage/injury was not identified. Thus, neuropsychological assessment as it is practiced today grew out of the need to clarify pathophysiological conditions where brain damage was not indicated by neurological, neuroradiological, or electrophysiological methods, in order to make differential diagnoses and provide information that would be useful in treatment planning and follow up (Dean & Gray, 1990). Based on the neuropsychological study of adults, this progressed to the application of neuropsychological methods and perspectives to the understanding of learning and other problems in children (L. C. Hartlage & Long, 1997).

Luria's (1970, 1980) theory, while based on adults, can be applied to children and adolescents. Neuropsychological techniques have been incorporated into the assessment of children for special education for some time (e.g., Haak, 1989; Hynd, 1981) with increasing interest by neuropsychologists in educational problems such as learning disability and attention deficit hyperactivity disorder (ADHD). The influence of theories specific to child psychology, school psychology, and education, are evident in the composition of neuropsychological assessment batteries, procedures, and measures used with children (Batchelor, 1996a). Increased interest and emphasis in the application of neuropsychology to educational issues and children may be due to a variety of factors: the emergence of clinical neuropsychology as a specialty area; advances in neuroscience and clinical evidence, specific to brain-behavior relationships, based on localized brain damage in childhood and youth; advances in technology (e.g., functional imaging) that are adding to the knowledge base regarding brain development and function; and continued research efforts specific to problems encountered by children and their neuropsychological functioning.

Several positive outcomes of the application of neuropsychology to children and adolescents have been identified. These include extending the range of diagnostic techniques available, providing for better integration of behavioral data (Dean, 1986; Gray & Dean, 1990; Obrzut & Hynd, 1983), increasing the ability to deal

adequately with the multidimensionality of observed behavior, creating a unified or holistic picture of a student's functioning (Rothlisberg & D'Amato, 1988), and providing documentation of changes in behavior and development (Hynd & Willis, 1988). Clinical child neuropsychology provides a theoretical framework for understanding identified patterns of strengths and weaknesses, the relationships between strengths and weaknesses, and the extent to which these patterns remain stable or are subject to change over the course of development (Fletcher & Taylor, 1984; Temple, 1997). Increased understanding of the child's strengths and weaknesses can potentially be used to identify areas that may provide difficulty for the child in the future, as well as compensatory strategies or methods to circumvent these difficulties. It has further been argued that neuropsychological assessment of children can provide a better understanding of the ways in which neurological conditions impact on behavior and the translation of this knowledge into educationally relevant information (Allen, 1989). Although not all psychologists agree with the application of neuropsychological principles to children (see Riccio, Hynd, & Cohen, 1993), the growing significance of clinical child neuropsychology is evident in the increasing number of child clinical and school psychology graduate programs that offer coursework in neuropsychology (e.g., D'Amato, Hammons, Termini, & Dean, 1992).

As a result of this growing interest in clinical child neuropsychology, the extent of knowledge available regarding the developing brain has increased dramatically since the 1980s. This includes advances in the understanding of typical development of neuropsychological functions (e.g., Ardila & Roselli, 1994; Halperin, McKay, Matier, & Sharma, 1994; Miller & Vernon, 1996; Molfese, 1995). Research has also begun to explore physiological processes and subcortical motivational systems that, together with environmental influences, are believed to impact on how the relevancy of information is determined and, ultimately, on the formation of cognitive representations in typically developing children (Derryberry & Reed, 1996). Additional advances in educational arenas have been made in the understanding of learning disabilities (e.g., Feagans, Short, & Meltzer, 1991; Geary, 1993; Obrzut & Hynd, 1983; Riccio, Gonzalez, & Hynd, 1994; Riccio & Hynd, 1995, 1996) as well as in the understanding of the short- and long-term problems associated with traumatic brain injury (e.g., Bigler, 1990; Snow & Hooper, 1994); the sequelae of neurological impairment of known causes such as lead poisoning, meningitis, and so on (e.g., Bellinger, 1995; Taylor, Barry, & Schatschneider, 1993); the

impact of cancer treatment on CNS function (e.g., Copeland et al., 1988); and the short- and long-term sequelae in children identified as at-risk for learning problems due to perinatal or prenatal difficulties (e.g., Breslau, Chilcoat, DelDotto, & Andreski, 1996; Cohen, Beckwith, Parmalee, & Sigman, 1996; Gatten, Arceneaux, Dean, & Anderson, 1994; Saigal, 1995; Waber & McCormick, 1995).

While more research has focused on educational problems, increased risk for psychiatric disorder and long-term adjustment problems have been found to be associated with brain injury, both in adults and children (e.g., Breslau & Marshall, 1985; Rutter, Graham, & Yule, 1970; Seidel, Chadwick, & Rutter, 1975). Research consistently demonstrates that adjustment and behavioral problems are associated with children who have neurodevelopmental deficits (e.g., Hooper & Tramontana, 1997; Tramontana & Hooper, 1997). Children with neurological impairment have been found to be six times more likely to develop emotional, behavioral, or motivational problems secondary to, if not as a direct result of, the neurological impairment (Dean, 1986). At one time it was believed that specific relationships between brain dysfunction and child psychopathology would be found; it is now posited that the relationships between brain integrity and psychopathology are nonspecific and impacted by secondary influences including failure, frustration, social stigma, family reaction, and so on (Tramontana & Hooper, 1997). Many advances have been made in the area of psychopathology, including the development of models specific to the underlying neurological basis of ADHD (see Riccio, Hynd, & Cohen, 1996), autism (e.g., Damasio & Maurer, 1978; Hooper, Boyd, Hynd, & Rubin, 1993; Hurd, 1996; Maurer & Damasio, 1982; Shields, Varley, Broks, & Simpson, 1996), schizophrenia in childhood and adolescence (e.g., Asarnow, Asamen, Granholm, & Sherman, 1994; Asarnow, Brown, & Strandburg, 1995; Hendren, Hodde-Vargas, Yeo, & Vargas, 1995), conduct disorder (e.g., Moffitt, 1993), and anxiety (e.g., Gray, 1982). Various models (e.g., Gray, 1982; Kinsbourne, 1989; Nussbaum et al., 1988; Rourke, 1989; Tucker, 1989) have been proposed to explain the interface between brain function and behaviors associated with childhood psychopathology.

This chapter will provide an overview of the neuropsychological assessment process for children, both historically and in the context of current practices and future trends. Continuing concerns with regard to the translation to children and adolescents of what is known about adult functioning and neuropsychological assessment, as well as concerns with measure-

ment and methodology, will be discussed. Finally, future directions and issues that need to be addressed in the neuropsychological assessment of children, if clinical child neuropsychology is to continue to add to the understanding of underlying processes in children's learning and behavior, as well as the application of that understanding to intervention programs, will be addressed.

4.10.1.1 Assessment Process

Neuropsychological assessment generally includes assessment of a number of functional domains that are, based on clinical evidence, associated with functional systems of the brain. This is considered important for the development of hypotheses and potential interventions (L. C. Hartlage & Telzrow, 1986; Whitten, D'Amato, & Chitooran, 1992). Areas evaluated generally include cognition, achievement, and behavior/personality/emotionality, as would be assessed as part of a general psychological evaluation. A neuropsychological evaluation provides for consideration of a wider array of functions, however, than is addressed in a typical psychological or psychoeducational evaluation (Dean, 1985, 1986; Obrzut, 1981). In general, the neuropsychological evaluation is more thorough and also includes the assessment of perceptual, motor, and sensory areas, and of attention, executive function (planning, organization), and learning/memory (e.g., Dean & Gray, 1990; Obrzut, 1981; Shurtleff, Fay, Abbot, & Berninger, 1988).

Given that the neurodevelopment of Luria's functional systems and the experiences of the child interact in a reciprocal manner (Spreen, Risser, & Edgell, 1995), as well as the potential for adjustment/behavioral difficulties, the use has been advocated of a transactional model that takes into consideration the reciprocal interactions of the child, home and family members, classroom (teacher and peers), and other social environments in which the child functions (Batchelor, 1996b; D'Amato & Rothlisberg, 1996; D'Amato, Rothlisberg, & Leu, in press; Teeter, 1997; Teeter & Semrud-Clikeman, 1997). This should incorporate information from a variety of sources (e.g., parents, teachers, physicians, medical records, school records, and so on) in order to enable cross-comparison (Batchelor, 1996b). In addition, it has been suggested that motivational factors (Batchelor, 1996b) and the child's ability to cope with the injury/impairment need to be determined (Dean, 1986). Thus, the neuropsychological assessment process not only incorporates a more complete review of information regarding the child but attempts to integrate this

information with an understanding of brain-behavior relations and environmental factors (Taylor & Fletcher, 1990).

This means that a neuropsychological assessment involves a wide range of tasks focused on the child, as well as measures/observations of the various contexts in which the child functions and the associated expectations. Some critics of neuropsychological assessment have argued that so extensive an evaluation is not time- or cost-effective (e.g., Little & Stavrou, 1993). For example, neuropsychological assessment of a learning disability goes beyond identifying the academic deficit(s) to the identification of the child's processing strengths and deficits as well as the child's ability to function in a variety of contexts (Morris, 1994). The assessment of a wider range of higher cortical functions is supported by research findings that neurological disorders are seldom expressed as a single dysfunction (Dean & Gray, 1990), and it has been shown to improve differential diagnosis of learning problems (D'Amato, Rothlisberg, & Rhodes, 1997; Morris, 1994; Rourke, 1994). It is further argued that the process of deriving hypotheses for intervention planning is a complex process that requires a comprehensive assessment battery coupled with neuropsychological foundations and familiarity with the contexts and task demands of the child (Gaddes, 1983; Rourke, 1994). The cumulative performances of the child on neuropsychological measures are seen as behavioral indicators of brain function (Fennell & Bauer 1997). Based on all of the data generated in the evaluation process, hypotheses are generated which are specific to how and why a child processes information (D'Amato, 1990; Dean 1986; Leu & D'Amato, 1994; Whitten et al., 1992). Inferences are then made based on the child's performance on a variety of measures and the theoretical perspective of the clinician.

Much of the skepticism regarding the application of neuropsychology to problems of learning and behavior in children has centered on the assessment-intervention interface (Little & Stavrou, 1993; Samuels, 1979; Sandoval & Halperin, 1981). It is argued by some that neuropsychological perspectives do not add to the ability to develop remedial and treatment programs and may even lead to a sense of hopelessness (Little & Stavrou, 1993). Others have argued that the information obtained from neuropsychological assessment can be used as the basis for developing appropriate intervention programs (Gaddes & Edgell, 1994; Reynolds, 1981b; Rourke, 1991, 1994; Teeter & Semrud-Clikeman, 1997). Data about the additional areas of functioning included in the neuropsychological assessment are needed to

support inferences about the integrity of various functional systems of the brain (Shurtleff et al., 1988). The neuropsychological perspective leads to better understanding of underlying causes of learning and behavior problems; this in turn results in an increased ability to develop appropriate interventions or circumvent future problems (D'Amato et al., 1997).

Ultimately, data generated from the neuropsychological assessment process are used to develop recommendations regarding whether the individual would profit from compensatory strategies, remedial instruction, or a combination of approaches (Gaddes & Edgell, 1994). Through the use of information about how various skills correlate in the developmental process, neuropsychological assessment allows one to make inferences not only about those skills measured, but also about skills that have not been evaluated. Further, by understanding the neurological correlates of these skills and of instructional methods, neuropsychological assessment can assist in the formulation of hypotheses regarding potential instructional methods/materials for a particular child (Reynolds, Kamphaus, Rosenthal, & Hiemenz, 1997). For example, based on the neuropsychological evaluation of two children with autism, differing nonverbal teaching strategies were identified for each child in order to improve their individual outcomes (Hurd, 1996).

While "treatment" within the framework of education is generally considered to consist of eligibility, placement decisions, and the development of an educational plan, "treatment" resulting from a neuropsychological assessment frequently includes assistance with specific medical management, vocationally related goals, speech/language areas, and physical issues (Cohen, Branch, Willis, Weyandt, & Hynd, 1992; Dean & Gray, 1990). Effective interventions need to take into consideration the myriad psychosocial contexts in which the child functions and adjustment and motivational issues, and to identify those environmental modifications that can ameliorate or reduce the behavioral effects of brain dysfunction (Batchelor, 1996b; Teeter & Semrud-Clikeman, 1997). It has been suggested that the interventions developed must also be multidimensional and incorporate not only academic, behavioral, and psychosocial techniques, but also include motivational, metacognitive, medical, and classroom management techniques (Batchelor, 1996b; Teeter & Semrud-Clikeman, 1997). However, correct diagnosis and early implementation of treatment strategies that work have been shown to be cost-effective in dollars and in the quality of a child's life (Reynolds, Wilen, & Stone, 1997).

4.10.2 MEASURES USED IN THE ASSESSMENT OF CHILDREN

As previously noted, the application of neuropsychological theory and assessment with children was derived from applications with adults. In the development of clinical child neuropsychology, historically, one of the basic avenues used in determining the assessment measures and processes to be used consisted of modifying, for use with children, existing neuropsychological batteries and other measures already used for adults (L. C. Hartlage & Long, 1997). In some cases, this involved modifying some tasks in the battery or adding tasks. An alternative strategy involved collecting some normative data on children for existing tasks. Both of these strategies were based on the clinical efficacy of the measures with adults, not with children, and on the assumption that tasks for adults measure the same thing when used with children. Similarly, in the assessment and hypothesis generation process, it is tempting to assume that neuropsychological findings from adults will be useful with children; however, this has not been shown to be a valid assumption.

When applied to children and adolescents, the premise that behavior can be used to make inferences about brain function and integrity has to be expanded to include consideration of neurodevelopmental differences that exist as a function of the age of the child. To directly apply adult inferences/hypotheses to children ignores what is known about changes in the functional organization of the brain as children grow (Cohen et al. 1992; Fletcher & Taylor, 1984). Research has provided evidence of age-based differences in children for verbal memory (Kail, 1984; Miller & Vernon, 1996), language (Segalowitz, 1983) and right hemisphere functions (Bakker, 1984; Wittelson, 1977). Recent research, for example, found that the relationship of memory, general intelligence, and speed of processing in children was not consistent with adult models (Miller & Vernon, 1996). Research has also suggested that typologies generated from the Halstead-Reitan Neuropsychological Battery (HRNB) used with children differed from typologies generated with adults, in that the child groups were more homogeneous, but provided less coverage (28–42%) as compared to adult typologies (Livingston et al., 1997). Because of neurodevelopmental changes, it is also not possible to view brain dysfunction as a continuum based on behavioral deficits as these may change over time (Fletcher & Taylor, 1984).

Further, there is often an over-reliance on signs of dysfunction in adults as reflecting pathology in children when these may be developmental. For example, on clock-face

drawing in adults, identification of unilateral spatial neglect has been of particular interest (e.g., Heilman, Watson, & Valenstein, 1985; Mesulam, 1985). Developmental study of clock-face drawing with children found that this hemispatial neglect was developmental and not infrequent through the age of seven years (Edmonds, Cohen, Riccio, Bacon, & Hynd, 1993). It was concluded that this developmental pattern was consistent with the development of the frontal lobes and planning ability in children.

Neurodevelopment follows an ontogenetic course with primary cortical zones generally mature by birth (Luria, 1980). Secondary and tertiary areas continue to develop postnatally. These include the integrative systems involved in the higher order functions of learning, memory, attention, emotion, cognition, and language as well as the association areas. The association areas are the last of these areas to develop and myelinate (Goldman & Lewis, 1978; Goldman-Rakic, 1987). Vygotsky (1980) suggested that not only is there continued development of secondary and tertiary areas, but that the interaction of primary, secondary, and tertiary areas is likely to change with chronological age (Merola & Leiderman, 1985; Rutter, 1981). Although the developmental sequence for the formation of neural pathways and the myelination of specific locations corresponding to specific behaviors have been identified, these do not correspond directly to models of cognitive development (Spren et al., 1995).

Knowledge of typical neurodevelopmental progress has increased since the 1980s; however, most of what is practiced today, as well as the theoretical bases in neuropsychology, is grounded on observations and informal assessment of individuals with identified brain damage (Reynolds, 1997b). Extensive research regarding typical neurodevelopment, particularly in relation to higher order cognitive skills, is limited, and the changing organization over time of brain function in children is only beginning to be understood (Hynd & Willis, 1988). Thus, there are still many unanswered questions regarding the developmental progression of many functional systems, particularly at the associative and integrative levels, and concerning how the neurodevelopmental progression maps onto the cognitive functioning observed.

It is often assumed, for example, based on earlier theory, that children reach adult levels of performance at 8–10 years of age. For example, Luria (1966) suggested that the frontal lobes become functional between the ages of four and seven years. This in turn led to the assumption that executive functioning would approach adult levels by age 8–10 years. It has been suggested that the greatest period of frontal lobe

development occurs at the six- and eight-year-old levels, which is consistent with Luria's initial hypothesis (Passler, Isaac, & Hynd, 1985). Subsequent research, however, has demonstrated that the development of frontal lobe functioning continues at least through age 12 and possibly through age 16 (e.g., Becker, Isaac, & Hynd, 1987; Chelune & Baer, 1986; Levin et al., 1991; Welsh, Pennington, & Grossier, 1991). Further, while cognitive ability does not appear to be a factor for particular measures of frontal lobe functioning after age 12, it has been suggested that cognition can impact performance on frontal lobe measures in younger children (Chelune & Thompson, 1987; Riccio, Hall, et al., 1994). Thus, it is important to first have a strong foundation of understanding of the normal neurodevelopmental course before it is possible to interpret accurately and differentiate behaviors that represent an alteration or deviance from expected neurodevelopment.

Not only do neurodevelopmental courses need to be considered, there are complex differences between children and adults in the mechanisms of brain pathology that lead to neuropsychological and behavioral/affective problems and these do not necessarily follow a similar progression in children as for adults (Fennell & Bauer, 1997; Fletcher & Taylor, 1984). The developing brain of the child needs to be considered in that the impact of neurological insult is influenced by age as well as location and nature of injury, gender, socioeconomic status, level of emotional adjustment and coping, and the individual's own adaptive skills (Bolter & Long, 1985). With the development of the child occurring on a continuous basis and at a rapid rate, it is often difficult to obtain sufficient consistency from the premorbid history (Batchelor, 1996a). Accurate estimation of premorbid ability levels is best obtained from previous individualized standardized cognitive or achievement assessment, or if this is unavailable, from results of group-administered standardized data from school records with some consideration for potential regression effects (Reynolds, 1997c). For young children, this information is not generally available. Prenatal and perinatal, as well as postnatal developmental histories may be inaccurate, incomplete, or unknown, particularly in very young children (Batchelor, Gray, Dean, & Lowery, 1988; Gray, Dean, & Rattan, 1987). Even in school-aged children, teacher reports, grades, and so on may result in inaccurate estimations of premorbid ability (Reynolds, 1997c).

Given the different mechanisms and progression involved in the pathology, it is clear that the inferences drawn from and the interpretations

of neuropsychological performance need to be different for adults and children. For children, the nature and persistence of learning problems is dependent on the status of development of various brain structures, the effects of the injury/insult, and the interactions between functional and dysfunctional neurological systems, as well as genetic and environmental influences (Teeter & Semrud-Clikeman, 1997).

Neuropsychological assessment of children and adolescents requires not only tests/measures that are age-appropriate and have sufficient empirical support for the inferences being made between neurological substrates and the behavioral performance of the child, but the generation of inferences also needs to take into consideration these developmental issues (Cohen et al., 1992). Further, it is important to document the sensitivity of the measures to neurobehavioral and neurodevelopmental functioning in children (Fletcher & Taylor, 1984). Although the measures are derived predominantly from neuropsychological study and clinical evidence regarding adults with known brain injury, a developmental perspective needs to be maintained in the application of neuropsychology to children (Hooper & Tramontana, 1997). Unfortunately, many of the measures used with adults do not have the sensitivity necessary to reflect developmental issues and, as a result, the utility of procedures used with adults in the neuropsychological assessment of children has multiple pitfalls and has been questioned (e.g., Cohen et al., 1992; Fletcher & Taylor, 1984).

4.10.2.1 Neuropsychological Interpretation of Children's Measures

Another approach to applying neuropsychological principles in the assessment of children took measures already in use for children (e.g., standardized intelligence tests) and interpreted these measures from a neuropsychological perspective; where existing child measures did not exist, these measures were then developed. L. C. Hartlage and Long (1997) indicated that most practitioners preferred this method (interpreting child-based measures from a neuropsychological perspective) as opposed to using adult measures with child norms. As with adults, this has occurred most frequently with the Wechsler scales. General summary scores of Wechsler scales have been found to be reliable indicators of brain integrity (Black, 1976; Hynd & Willis, 1988). Various subtests of the WISC-R also have been found to correlate with neuropsychological measures (see Batchelor, Sowles, Dean, & Fischer, 1991) and have been used to

formulate hypotheses (Kaplan, 1988). Multiple efforts have been made with regard to recategorizing or clustering various subtests to provide for neuropsychological interpretation of the WISC-R. L. C. Hartlage (1982), for example, suggested that the functional integrity of the right and left hemispheres could be estimated by comparing the Similarities and Picture Arrangement subtests (temporal lobe) and the Arithmetic and Block Design subtests (parietal lobe). Bannatyne (1974) proposed four categories of neuropsychological function that could be assessed and interpreted based on combinations of subtests on the WISC-R: verbal comprehension, sequencing, spatial, and acquired knowledge. Kaufman (1979) recategorized the subtests into successive and simultaneous tests, based on Luria's theory.

Concerns with this practice have been evidenced in the literature. Interpretations based on isolated measures of a child's behavior (e.g., a single subtest) have limited reliability and validity (Kamphaus, 1993; Lezak, 1995) and this is often what occurs in this process. Recategorizations of multiple subtests (e.g., Bannatyne, 1974; Kaufman, 1979), appear to have greater reliability, but the validity of these recategorizations continues to be questionable (see Kamphaus, 1993). Further, in many cases there is no attempt to translate the inferences made, using these methods, into effective interventions.

4.10.2.2 Development of New Measures for Children

As opposed to trying to "make do" with existing children's measures, additional measures have been developed with an underlying neuropsychological basis. For example, the Luria-Das model of successive/simultaneous processing (Das, Kirby, & Jarman, 1979) in conjunction with the cerebral lateralization research by Sperry (1968, 1974), Kinsbourne (1975), and others, served as the basis for the development of the Kaufman Assessment Battery for Children (KABC; Kaufman & Kaufman, 1983a). As such, the design of the KABC is compatible with current neuropsychological models of higher order cognitive function (Reynolds & Kamphaus, 1997). Unlike the Wechsler scales, where mode of presentation determines the scale with which a task is associated, on the KABC the cognitive processing demands of the task (e.g., simultaneous or sequential) determine the scale with which it is associated (Kaufman & Kaufman, 1983b). Further, rather than conceptualizing lateralization based on content or method of presenta-

tion, the lateralization component of the KABC is based on the way in which the information is processed or manipulated. Within each scale, there is a variation of mode of presentation and response that allows for further evaluation of complex functional systems (Reynolds & Kamphaus, 1997). KABC interpretation is intended to identify cognitive neuropsychological strengths of the child, and the related instructional methods and learning activities that will exploit these strengths and circumvent deficit areas. Research on the effectiveness of this model for intervention is, however, limited.

Evaluation of the KABC with regard to its relevance to Luria's approach and to child neuropsychology has been positive (e.g., Donders, 1992; Majovski, 1984; Snyder, Lark, Golden, Grove, & Allison, 1983). It has been suggested that the KABC is a good complement to other neuropsychological tests. Specifically with regard to the use of the KABC as a component of a neuropsychological battery, it has been shown to provide useful information in the differential diagnosis of learning disability subtypes (e.g., Hooper & Hynd, 1985; Telzrow, Century, Harris, & Redmond, 1985) and right hemisphere dysfunction, which is consistent with physical evidence (Morris & Bigler, 1985; Shapiro & Dotan, 1985). Similar positive results were found in the comparison of dichotic listening performance and KABC results (Dietzen, 1986). Thus, the KABC has been shown to be sensitive to traumatic brain injury to specific cortical regions (Donders, 1992). Research also indicated that the pathognomonic and intellect scales of the Luria Nebraska Neuropsychological Battery-Children's Revision were closely related to performance on the global scales of the KABC (Lark, Snyder, Grove, & Golden, 1983). Research results overall tend to support the use of the KABC in neuropsychological assessment, and subtests of the KABC are frequently used in eclectic batteries (e.g., Nussbaum et al., 1988; Branch, Cohen, & Hynd, 1995).

The KABC may well be the test of choice for children under age five (Reynolds et al., 1997); the use of sample and teaching items adds to the likelihood that a neurological substrate or functional system is being assessed as opposed to language, experience, or culture (Reynolds & Kamphaus, 1997). The KABC has strong validity and reliability (Kamphaus, 1993), is sensitive to developmental changes in information processing/functional organization (Reynolds & Kamphaus, 1997), and is considered an appropriate instrument for use with US ethnic minorities (e.g., Fan, Willson, & Reynolds, 1995; Kamphaus & Reynolds, 1987). While further research with the KABC in conjunction with neuropsychological assessment is needed,

available research supports the potential for the KABC to be a useful tool for child neuropsychologists with results providing implications for the habilitation of learning problems (Reynolds & Kamphaus, 1997).

4.10.2.3 Current and Future Trends

The development of new measures, specifically designed and normed for children may not only reflect current interest areas in children's learning and behavior, but may in many ways dictate the future directions of neuropsychological assessment of children. In particular, since the late 1980s a number of measures have been developed which are specific to memory and attention. At the same time, there is also an increase in the use of technology, with or without the inclusion of electrophysiological or imaging methods, which is evident in the research literature and clinical practice.

4.10.2.3.1 Memory

Nearly every disorder that involves the CNS and higher cognitive functions includes some form of memory complaint; memory is incorporated in almost all daily activities (Reynolds & Bigler, 1997a). Research across neurological disorders points to the importance of memory in evaluating brain integrity (Reynolds & Bigler, 1997a); 80% of a sample of clinicians who performed testing noted memory as important (Snyderman & Rothman, 1987). Standard psychoeducational batteries used with children tend to focus solely on cognitive ability as defined by IQ, achievement, and behavioral status. In the area of learning disabilities, there has been recent interest in examining the underlying psychological processes, and particularly learning and memory (Zurcher, 1995). Research in the area of memory and the development of new measures to assess memory functions may lead to further interest in the learning process itself (Reynolds, 1992). It has been argued, additionally, that the assessment of learning and memory would provide useful information for instructional planning (Wasserman, 1995).

Historically, assessment of memory in children relied on the use of subtests from various tests including the KABC, the WISC-III and its earlier versions, and so on (e.g., Nussbaum et al., 1988; Branch et al., 1995). All too frequently, inferences regarding verbal memory in particular relied on the Digit Span subtest of the Wechsler scales. Multiple concerns about relying on Digit Span can be found in the research literature (e.g., Reynolds, 1997a; Talley, 1986). Recent research in the area of memory has

suggested that the traditional combining of forward and backward digits may be inappropriate and that these tasks represent quite different cognitive demands (Ramsey & Reynolds, 1995; Reynolds, 1997a) with distinct neuropsychological substrates. Initial findings suggest, for example, that forward memory span may be more directly impacted by attention while backward memory span may be more a reflection of general intelligence. Additional investigation into the distinction between forward and backward memory span, as well as into other areas of memory continues to be needed. Due to the increased interest in this area, children's norms for measures used in the assessment of memory in adults have been developed (e.g., Delis, Kramer, Kaplan, & Ober, 1994). In addition, three comprehensive measures for the assessment of memory/learning have been developed specifically for children and adolescents since the mid-1980s. The development of these measures has in many ways been due to the perceived inappropriateness of adult measures of memory for use with children and the inability to relate results from adult measures to the contexts (e.g., school) in which children function.

The first of the measures developed for the assessment of memory in children, the Wide Range Assessment of Memory and Learning (WRAML; Sheslow & Adams, 1990), consists of 12 subtests which yield verbal memory and visual memory scores, with normative data for children ages 5–17 years. Delayed recall trials can be given for four of the subtests. Initial factor analysis of the WRAML corroborated the two-factor structure (Haut, Haut, Callahan, & Franzen, 1992); however, with at-risk children and a clinical population, three factors were extracted (Aylward, Gioia, Verhulst, & Bell, 1995; Phelps, 1995). Some concern has been voiced with regard to the multiple items/tasks that may tap attention as opposed to memory and the absence of consideration of attention/concentration (Haut et al., 1992). Further, evaluation of the WRAML for children with, compared with those without, ADHD or learning disabilities indicated that the WRAML provided little additional information for discriminating between clinical groups (Phelps, 1996).

The Test of Memory and Learning (TOMAL; Reynolds & Bigler, 1994) consists of 10 core subtests (five verbal and five nonverbal) yielding separate verbal memory and nonverbal memory scale scores as well as a composite memory score. A delayed recall procedure can be implemented to provide a delayed recall index. Additional supplemental indices (e.g., sequential recall, free recall, attention/concentration,

and learning) can also be computed. Using a variety of factor analytic methods, Reynolds and Bigler (1996) examined the latent structure of the TOMAL. Factor analytic study of the TOMAL indicated that the factor solutions obtained were highly stable across all age groups. Notably, none of the solutions obtained matched the verbal-nonverbal dichotomy usually considered and represented by the two scales of the TOMAL. Instead, what emerged were components representing various levels of complexity in memory tasks and processing demands that cut across modalities. Alternative methods of interpretation based on the factor analytic results are available (see Reynolds & Bigler, 1996). The TOMAL does provide separate scores for forward and backward recall, in contrast to many scales that combine these inappropriately. Unlike most neuropsychological measures (Reynolds, 1997b), the TOMAL included studies of ethnic and gender bias during standardization; items showing cultural biases were eliminated.

Most recently, the Children's Memory Scale (CMS; Cohen, 1997) was developed with linkages to the WISC-III built in to the standardization process. The composition of the CMS was based on extensive clinical practice with initial tasks and items, field trials of the measures, and feedback from clinicians involved in the field trials. The CMS consists of six core subtests representing verbal memory, attention/concentration, and visual/nonverbal memory as well as three supplemental subtests. The CMS provides for evaluation of immediate recall as well as delayed recall of the verbal and nonverbal memory areas. For scoring purposes, seven index scores can be calculated to examine differences between immediate/delayed verbal/visual memory, learning, recognition, and attention/concentration. Factor analytic studies of the standardization sample were conducted and four models evaluated to determine the "best fit." Results indicated that the three-factor solution (attention/concentration, verbal memory, visual memory) was the most consistent (Cohen, 1997).

4.10.2.3.2 Attention

It has been argued that the most frequent symptoms associated with childhood neuropsychological disorder include attention/concentration, self-regulation and emotional/behavioral problems (Nussbaum & Bigler, 1990). Further, it is the neural traces left by attention that are likely the root of memory. It is not surprising that there is increased interest in the measurement of attentional processes or that these are seen as an important component

of the assessment process. The assessment of attention, more so than of other domains, has moved to computerized approaches. The most comprehensive battery of computerized measures is the Gordon Diagnostic System (GDS; Gordon, 1983). This is a microcomputer-based assessment that includes 11 tasks specific to attention and self-regulation. Since the development of the GDS, a number of other computer-based measures of attention and impulsivity have been developed and marketed. These programs tend to vary with regard to the actual paradigm used; there are variations in the modality employed, the type of stimuli, and the nature of the task (Halperin, 1991).

Continuous performance tests (CPTs), for example, may require a response only when a specified target stimulus is presented (if X) or only when the target stimulus follows another specified stimulus (if AX) and so on. A further variation of this is a similar task where the required "response" to the presentation of the target stimuli is, however, to inhibit responding (Conners, 1995). The stimuli may be presented in a visual or auditory format, or in a combination format requiring a modality shift. Also, depending on the program used, the scores may be limited to correct responses, commission errors, and omission errors, or may include reaction time information.

Through the use of computerized measures of attention, knowledge specific to the developmental nature of attentional processes has been gleaned (Mitchell, Chavez, Baker, Guzman, & Azen, 1990). Research has demonstrated the usefulness of computerized measures of attention and self-regulation for monitoring the effects of medical management (e.g., Barkley, DuPaul, & McMurray, 1991; Barkley, Fischer, Newby, & Breen, 1988; Hall & Kataria, 1992). It was anticipated that computerized assessment of attention would provide more objective data in the assessment process for ADHD as well as providing information specific to attentional deficits associated with traumatic brain injury or other neurological disorders (Timmermans & Christensen, 1991). The results of studies with various paradigms for CPTs are equivocal with regard to discriminant validity specific to ADHD (e.g., Barkley et al., 1991; Wherry et al., 1993) as well as concerning the extent to which results are consistent with teacher perceptions (Barkley, 1991, 1994). Interpretation of these measures is limited by the availability of comprehensive research with any one software program. The extent to which cultural differences, gender differences, cognitive ability, order effects, and so on impact on CPT performance is unknown. Further, the extent to which the particular paradigms used

provide predictive information that may be helpful in intervention planning has not been studied.

4.10.2.3.3 Computer-administered assessment

Measures of attention are not the only computer-based assessment tools. A computerized neuropsychological test battery for adults has been developed (Powell et al., 1993), computer-administered interviews and self-reports are available, and specific neuropsychological tests or their analogs can be administered via computer (e.g., Burin, Prieto, & Delgado, 1995; Heaton, 1981). Computerized assessment of children's reading skills has been investigated with indications of high coefficients of equivalence with traditional assessment (Evans, Tannehill, & Martin, 1995). With advances in microcomputers, the use of computerized assessment will likely increase in the near future.

The use of computers and technology in assessment has a number of advantages and clearly allows for the development of an increasing variety of tasks without excessive and cumbersome testing materials; computerized assessment may be less time-consuming and, as such, cost- and time-effective. Further, the speed or measure of time to task completion is considered one of the most sensitive indices in neuropsychological assessment and computer programs can provide increased accuracy in the measurement of speed of processing (Kane & Kay, 1997). Kane and Kay (1997) point out a number of additional advantages to the use of computers in the assessment process, including presentation of items at a fixed rate (computer-paced) as well as providing for accurate measure of time to completion (child-paced). Computers can also be used to generate multiple forms of a test, thus providing baseline data as well as a means of monitoring change over time. With computerized assessment, standard/uniform administration is ensured and results are free of potential bias. Computers further facilitate the production of relevant test statistics (Kane & Kay, 1997).

There are however, multiple concerns and disadvantages with "diagnosis by computer." Predominant among these is the loss of information from not being able to observe the process and strategy used by the individual in reaching the solution (Powell, 1997). First (1994) concluded that computerized assessment processes were advantageous, but cautioned that clinicians must continue to be a strong component in the diagnostic process in order to provide for diagnostic validity. Computers cannot replace the information gained from interaction and clinical observation of process

nor can a computer draw conclusions regarding level of attention, motivation, fatigue, and so on that may be cues to discontinue testing for a brief period. Computers also cannot provide the child with prompts and encouragement as needed to maintain performance over time (Kane & Kay, 1997). At the extreme, there is the potential for computers to be used as a substitute for a complete evaluation and this is of concern (Kane & Kay, 1997). First (1994) asserted that clinicians needed to be well advised of the limitations as well as the strengths of computerized assessment procedures. As the number of computer-driven assessments increases, there will need to be an analogous increase in the research field comparing the various programs and their psychometric properties with each other and with more traditional methods of assessments. At the time of writing, in the late 1990s, many computerized assessment methods fail to meet established testing standards (Kane & Kay, 1997).

4.10.2.3.4 Integration of neuroimaging and electrophysiology

With advances in neuroscience, clinical and research protocols may more frequently include neuroradiological methods in conjunction with neuropsychological techniques in order to enhance understanding of childhood disorders. This type of "partnership" is already occurring in a number of research areas (e.g., Bigler, 1991; Denckla, LeMay, & Chapman, 1985; Duffy, Denckla, McAnulty, & Holmes, 1988; Hynd, Marshall, & Semrud-Clikeman, 1991). The integration of information from neuroradiology with neuropsychological assessment has already established relationships for specific lesions and associated behaviors and is beginning to establish a better understanding of the relationship between myelination differences and white/gray matter ratios (e.g., Harbord et al., 1990; Jernigan & Tallal, 1990; Turkheimer, Yeo, Jones, & Bigler, 1990). The availability of imaging using ultrasound has added to the knowledge of relationships between gross abnormalities evident *in vitro* and later negative outcomes (e.g., Beverley, Smith, Beesley, Jones, & Rhodes, 1990; Iivaneihan, Launes, Pihko, Nikkinen, & Lindroth, 1990). Measurement issues in imaging, such as differences in resolution from one magnetic resonance image to another, continue to be problems in this area, but will hopefully be resolved in the future. While in the past routine EEG of children did not offer much utility in the evaluation of learning or behavior problems, the development of computer-assisted analysis has improved the interpretability of electrophysiological measures (Duffy & McAnulty, 1990).

Computerized measures have been developed to examine more closely the speed of information processing, through reaction time paradigms that have included linguistic (e.g., Lovrich, Cheng, & Velting, 1996) as well as visual stimuli (Novak, Solanto, & Abikoff, 1995), in conjunction with electrophysiological measures.

This integration of methods across neuroscience and neuropsychology is providing further evidence concerning brain-behavior relationships and adding to the knowledge base related to neurodevelopmental processes in children and adolescents. Functional imaging and other imaging quantification methods hold promise for furthering the future understanding of neuropsychological performance (Bigler, 1996). Similarly, it has also been argued that a comprehensive and integrative assessment process, that involves both the neurologist and neuropsychologist with the tools and expertise of both disciplines, may enhance the value and role of neuropsychological assessment (Batchelor, 1996a, 1996b).

4.10.2.3.5 Integration of cognitive and developmental psychology

Neuropsychological assessment of children is being influenced more and more by developmental and cognitive psychology. This is most apparent in the areas of language, attentional processes, and executive functions (Williams & Boll, 1997). Integration across fields has been suggested specifically with regard to metacognition (from cognitive psychology) and executive function (Torgesen, 1994). The domain of "executive function" may incorporate a variety of constructs (e.g., attention, self-regulation, working memory) but the "executive" processes generally focus more on effortful and flexible organization, strategic planning, and proactive reasoning (Denckla, 1994). Denckla further asserted that executive function cannot be dealt with as a "composite" of scores on various measures, but must be fractionated. The measurement of executive function in children is exceptionally difficult due to the ongoing development and maturation of the frontal lobes through adolescence. Factor analytic study of executive function tasks (Welsh et al., 1991) yielded factors that appeared to be divided according to developmentally related constructs as opposed to theoretical ones (Denckla, 1994). The majority of measures for executive function which are used with children are downward extensions of adult measures and many lack sufficient normative data and psychometric study. In addition, the emphasis on the use of novel tasks in the assessment of executive function places significant limitations

on the evaluation of interventions in the area of executive processes. Torgesen (1994) argued that current measures of executive function, with the presumed assumption for a need for novelty, evidence a lack of cross-theoretical integration between neuropsychology and the information-processing paradigms. He further stated that there is a need to include assessment of tasks that are ecologically based and require executive function, in order to enhance the evaluation of treatment programs designed to remediate executive processes. Certainly, the production of child-centered, developmentally sensitive measures of executive processing, that are more directly linked to real-life activities thus facilitating the development of interventions, and that have sufficient flexibility to allow for pre- and postevaluation, is needed. Overall, integration of neuropsychological assessment and models of cognitive development may lead not only to a better understanding of deficit processes but also to better remediation/habilitation programs (Williams & Boll, 1997).

4.10.2.4 Measurement Issues

Although research methods and statistical tools have greatly improved since the early 1970s, clinical child neuropsychology has been criticized for its failure to attend to principles of research and to incorporate psychometric advances (Cicchetti, 1994; Parsons, & Prigatano, 1978; Reschly & Gresham, 1989; Ris & Noll, 1994; Sandoval, 1981; Willson & Reynolds, 1982). Problems with statistical methods and design in clinical neuropsychology have been frequently noted (e.g., Adams, 1985; Dean, 1985; Reynolds, 1986a, 1986b, 1997b).

One major concern relates to the extent and nature of normative data for many measures used in the neuropsychological assessment of children. Although clinical insight may be gained by observation of test performance, sound normative data provides a backdrop against which to evaluate that insight (Reynolds, 1997b). The systematic development and presentation of normative data across the lifespan for many tools used in neuropsychological assessment have received far too little attention to date (Reynolds, 1986b) and greater attention in this area is needed. Good normative data require extensive systematic and stratified sampling of a population in order to obtain a reliable standard against which to judge the performance of others. The provision of adequate normative data has multiple benefits for the field of neuropsychology, including improved communication among clinicians and researchers, increased accuracy in diagnosis,

and facilitation of training for new members to the discipline. In addition, good normative data provide the opportunity to deflate and expose a variety of clinical myths (Reynolds, 1986b).

Most of what is known about the measures used is specific to the performance of those with identified brain injury/insult as opposed to typically developing individuals. Normative data that are available in the literature are often based on small samples, may have been collected in a single geographical region, and do not reflect the ethnic diversity, socioeconomic levels, or gender composition of the general population. In many cases, the sample is predominantly male and Caucasian, yet research suggests that gender and cultural differences may also contribute to variations in brain organization (e.g., McGlone & Davidson, 1973).

The lack of sufficiently large, stratified samples in the development and standardization of neuropsychological assessment inhibits the understanding of demographic influences (Reynolds, 1997b) and thus complicates test interpretation (Dean, 1985). That cultural differences exist on standardized measures is well documented. Mostly, the use of neuropsychological measures with Hispanic populations has been studied (e.g., Ardila & Roselli, 1994; Ardila, Roselli, & Putente, 1994; Arnold, Montgomery, Castenada, & Langoria, 1994), but overall research on the effects of cultural differences (e.g., differences in the value of speed of responding) is sparse. Differences between ethnic groups have also been examined with respect to specific measures of memory (e.g., Mayfield & Reynolds, 1997). However, for most neuropsychological measures there has been no study of ethnic and gender differences; cultural/ethnic differences are infrequently accounted for in the collection of normative data and therefore cannot be used in the interpretation process.

All too frequently, neuropsychologists rely on the "clinical" nature of the test and overlook the psychometric concepts of reliability and validity. The need for the establishment of reliability of neuropsychological measures has been cited in the literature (e.g., Parsons & Prigatano, 1978, Reynolds, 1982); reliability information on neuropsychological measures is not routinely reported in research studies and is frequently not included in the test manuals (Reynolds, 1997b). Reliability of test scores is important as it relates to the amount of variance that is "real," systematic, and related to true differences between individuals. Therefore, it is important to determine the reliability of neuropsychological measures for purposes of individual diagnosis as well as for research, in that reliability influences the likelihood that any experimental or treatment effects will be

detected (Reynolds, 1986b). Reliability is also the foundation on which validity is built.

Related to issues of reliability and validity, the method of scaling/measurement used with any test or measure is "crucial" (Reynolds, 1997b, p. 189). Scaling across neuropsychological measures, however, is inconsistent. Frequently what are obtained are raw scores for number correct, time for completion, or number of errors. This results in the need to use score transformations, based on insufficient normative data, in order to make any kind of meaningful comparison. Alternatively, clinicians may use inappropriate scales such as age or grade equivalents in an attempt to give meaning to raw data (Reynolds, 1997b). Grade equivalents, in particular, are inappropriate due to the extent of extrapolation that is used in their derivation as well as faulty assumptions that are made with regard to learning and growth over time (e.g., from lower to upper grades, across subject areas, and throughout the calendar year). Further, grade equivalents exaggerate small differences in performance between individuals and for a single individual across tests (Reynolds, 1986b). It is, therefore, imperative that standard score conversions, based on adequate normative data, be provided for measures used in neuropsychological assessment.

It has been asserted that by age 10 or 12 children perform at adult levels in some areas, and for many older neuropsychological measures most of the normative sampling, in addition to using small numbers, often stopped at age 12. This is despite the fact that many researchers have suggested that neurodevelopment continues through at least age 14 (Boll, 1974) and possibly through age 16 (Golden, 1981). This further limitation in the provision of normative data impedes the interpretation process for adolescents, bolsters the assumption that adolescents should function as adults, and promotes the use of downward extensions of adult measures that often are not appropriate. Neuropsychological function is developmental, and distinct age-related norms are required. It has also been recommended that item response theory (IRT) be used to ensure that neuropsychological measures include an adequate range of difficulty levels, thus ensuring coverage of developmental levels (Morris, 1994). Most existing neuropsychological measures, however, have not been subjected to this type of analysis.

The standardization of administration procedures is also an area of concern. Reynolds (1986b) commented on the availability of at least four versions of Halstead's category test, three of which were somewhat similar and the fourth with significant differences in terms of administration. Despite these differences, however, the

same normative data are used. Similarly, administration of the Wisconsin Card Sorting Test (Heaton, 1981) can be done traditionally or via computer, yet there is a single normative data set to be used for scoring and interpretation. Differences in administration impact on the validity and reliability of the measure and normative data, including validity studies, for each variation of administration (unless controlled in the standardization process) are necessary.

Sensitivity, specificity, and diagnostic accuracy need to be further researched as well. Sensitivity is the extent to which a given test accurately predicts brain impairment and is often gauged by statistical power (Pedhazur, 1973). Sensitivity is dependent on validity. No single neuropsychological measure demonstrates high sensitivity (Boll, 1978); combined scores from a given battery may be more successful (e.g., Selz & Reitan, 1979a). In contrast, specificity is dependent on the nature of the behavioral, cognitive, and emotional functions of the task (Batchelor, 1996b). The extent of specificity can only be determined by comparing clinical groups to each other as opposed to focusing on differences between a specific clinical group and the normal population. Cross-clinical group comparisons are frequently not done however. When research is based on comparisons across clinical groups, the results are generally inconclusive (Koriat, Gualtieri, van Bourgondien, Quade, & Werry, 1985).

In comparing clinical groups, it is important to control for comorbidity and family history (Seidman et al., 1995) as well as to differentiate between subtypes of a given disorder when these have been identified (e.g., Halperin, 1991). For many disorders, subtypes have been validated, yet frequently research studies with clinical groups rely on the more global rubric. With regard to learning disabilities, Rourke (1994) asserted that this "lumping" together may lead to gross misunderstanding, if not to conflicting results across studies. For example, the need to develop and incorporate typologies/subtypes for homogeneous grouping of children with dyslexia, for the purposes of developing appropriate intervention as well as for research purposes, has been recognized for some time, yet much of the educational and psychological literature and practices relating to dyslexia continue to address heterogeneous groups of children without regard for subtype (Reynolds, 1986b). Differing typologies and comorbidity have rarely been considered in the extant literature on many neuropsychological measures and likely contribute to the conflicting results of differing studies. Further, the constructs being measured by given tasks may be

influenced by gender, premorbid status, the task itself, neuropsychological functions, and so on, making a high level of specificity difficult to attain (Batchelor, 1996b). Batchelor (1996b) suggested that many neuropsychologists compromise between sensitivity and specificity through the selection, administration, and interpretations of neuropsychological measures that are needed to effect such a balance.

Often, in an attempt to provide accurate differential diagnosis, a large set of behaviors is typically assessed. Researchers then use multivariate classification procedures for determination of group information or to determine the effectiveness of specific measures in the diagnostic process. The sample sizes in many of the studies, however, are too small for multivariate analysis, given the large number of variables involved. As a result, in the absence of cross-validation, many diagnoses or classifications may be due to random relationships (Willson & Reynolds, 1982). Problems with the lack of consistency in the diagnosis/classification of disorders also impede the research process and ultimately, clinical practice (Hooper & Tramontana, 1997).

4.10.2.5 Approaches to Test Selection with Children

In addition to selecting tests based on psychometric properties, it has been suggested that child neuropsychologists should select measures that vary along a continuum of difficulty, include both rote and novel tasks, and vary the tasks with regard to processing and response requirements within modalities (Rourke, 1994). Many neuropsychologists continue to include observation and informal assessment; others have adopted more actuarial approaches; many use a combination of observation, informal assessment, and actuarial approaches (Reynolds, 1997b), with a focus on direct appraisal of functions and abilities in order to obtain detailed information on the behavioral effects of brain impairment (Tramontana, 1983). It has been argued that the use of actuarial methods that rely on standardized measures to obtain information may not, however, be useful in intervention planning (D'Amato, 1990).

In addition to quantitative measures, neuropsychological assessment may incorporate not only Luria's theory but also his qualitative assessment model (Luria, 1966, 1970). Luria described assessment that was flexible and varied from individual to individual depending on the functional system that was of concern (Teeter, 1986). Although more dependent on clinician interpretation, qualitative methods can

add to information related to the process of learning and may be better suited to the development of intervention/treatment plans (D'Amato et al., in press). In the incorporation of qualitative tasks, child neuropsychologists make use of work samples, informal tasks, criterion-referenced measures and clinical observations of interactions throughout the assessment process (D'Amato et al., 1997). Qualitative procedures can also be used to complete a task analysis and determine specifically which components of a more complex task are problematic for the child (Taylor, 1988). Others may use standardized measures but administer them in other than standardized fashion (Kaplan, 1988). Modifications of tasks presented (e.g., provision of cues, adjustment of rate, changing modality of presentation or response, adjustment of task complexity) can provide insight into processing differences (Clark & Hostettler, 1995; Harrington, 1990; Ylvisaker et al., 1990) and have been recommended for use in the evaluation of children who are culturally or linguistically diverse (Gonzalez, Brusca-Vega, & Yawkey, 1997). Unfortunately, tests administered with these types of modifications are no longer consistent with standardization procedures, and clinicians need to exercise caution in the interpretation of brain-behavior relations based on qualitative data (D'Amato et al., in press). A strictly qualitative approach using experimental/ad hoc measures and nonquantitative/non-standardized interpretation of standardized measures may provide additional information, but does not allow for verification of diagnostic accuracy, is not easily replicated, and does not allow for formal evaluation of treatment methods (Rourke, 1994). In practice, most clinicians prefer a combination of quantitative and qualitative measures (Rourke, 1994).

Test selection in the neuropsychological assessment of children and young people varies considerably from clinician to clinician due to differences in philosophy and theoretical foundations. Evaluation may take the form of various published battery approaches (e.g., Golden, 1997; Reitan, 1974; Selz, 1981) or may use a more eclectic approach (e.g., Benton, Hamsher, Varney, & Spreen, 1983; Gaddes, 1980; Hynd & Cohen, 1983; Knights & Norwood, 1979; Obrzut, 1981; Obrzut & Hynd, 1986; Rourke, Bakker, Fisk, & Strang, 1983; Rutter, 1983; Spreen & Gaddes, 1979; Teeter, 1986; Tramon-tana & Hooper, 1987). Generally, however, the approaches can be categorized as nomothetic, idiographic, or a combination of these two approaches. Additional variation within categories, however, is evident in the extent to which clinicians rely on quantitative, qualitative, or both types of information in the process.

4.10.2.5.1 Nomothetic approaches

The fixed/standardized battery or nomothetic approach uses the same assessment protocol for all children being assessed. An example of the nomothetic approach would be the administration of a published neuropsychological battery, usually in conjunction with IQ and achievement tests. It may also be a predetermined set grouping of selected tests that remains constant across children evaluated, regardless of the referral problem (Sweet, Moberg, & Westergaard, 1996). These tend to be more actuarial in nature (Lezak, 1995) and often rely on cut-off scores, pathognomonic signs, or a combination, for determination of the presence of brain damage. The choice of a fixed battery approach is generally related to an orientation and preference consistent with standardized procedures, objective methods, and psychometric development. It may also reflect a preference for "blind" assessment such that the referral problem does not dictate the measures used (Goldstein, 1997). This approach has the advantage of covering a breadth/depth of functions, provides for extensive databases, and facilitates the collection of data for clinical interpretation of large numbers of clinical groups. Standardized/nomothetic batteries, however, often do not take into consideration education, age, and experiential variables, and may or may not specifically address the referral question. Further, diagnosis with a nomothetic approach may be driven by the base rates of the clinical problems in a particular setting, due to sampling bias, and therefore may not be useful for detecting disorders in other population samples (Tramontana & Hooper, 1987). Use of a standardized battery/nomothetic approach appears to be declining in the general area of clinical neuropsychology (Sweet et al., 1996); however it may be the preferred method if litigation is a potential issue (Reitan & Wolfson, 1985).

The published neuropsychological batteries most frequently used with school-aged children are the Luria Nebraska Neuropsychological Battery-Children's Revision (LNNB-CR; Golden, 1984), the Halstead-Reitan Neuropsychology Battery (HRNB; Reitan & Davison, 1974) and the Reitan Indiana Neuropsychological Test Battery for Children (RINB; Reitan, 1969). All of these batteries require extensive training for appropriate administration and interpretation of results. The neuropsychological battery is often supplemented with a traditional test of cognitive ability as well as achievement testing.

The HRNB and RINB are considered to be the most widely used in clinical practice (Howieson & Lezak, 1992; Nussbaum & Bigler,

1997). Both of these use a multiple inferential approach to interpretation, including level of performance, pathognomonic signs, patterns of performance, and right-left differences (Reitan, 1986, 1987). The batteries contain numerous measures that are considered necessary for understanding brain-behavior relationships in children and adolescents. Descriptions of these measures are provided in Tables 1 and 2. Both the HRNB and the RINB can be used in clinical practice for the assessment of a child with identified brain damage as well as with those children where specific brain damage has not been documented through neuroradiological methods (Nussbaum & Bigler, 1997).

Strong correlations have been found between the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) and the RINB and HRNB (Klesges, 1983) suggesting the ability of the latter tests to predict neuropsychological dysfunction. A number of factor analytic studies have been completed comparing results from the WISC-R and the Reitan batteries (e.g., Batchelor et al., 1991; D'Amato, Gray, & Dean, 1988; Snow & Hynd, 1985a); these consistently suggest that most of the variance is due to factors of language, academic achievement, and visual spatial skills. With the addition of other measures, up to eight factors were found (D'Amato et al., 1988; Batchelor et al., 1991). Although factor analytic research has demonstrated a good deal of common information when the WISC-R and HRNB were both given, it has also been determined that the HRNB offers unique information (Klonoff & Low, 1974) and the addition of the HRNB to the typical psychoeducational battery has been found to increase the extent to which variability in school achievement can be accounted for (Strom, Gray, Dean, & Fischer, 1987). Research regarding the efficacy of the HRNB in the differential diagnosis of children with learning problems is equivocal (Arffa, Fitzhugh-Bell, & Black, 1989; Batchelor, Kixmiller, & Dean, 1990; Selz & Reitan, 1979a, 1979b). Factor analytic research with the RINB has been less conclusive (Crockett, Klonoff, & Bjerring, 1969; Foxcroft, 1989; Teeter, 1986). The RINB has been found to be sensitive to mild levels of traumatic brain injury within four months of injury in the absence of obvious lags in academic achievement (Gulbrandson, 1984). The HRNB and RINB are both downward extensions of the adult version with some modifications for children (Teeter, 1986) and do not fully reflect the developmental continuum of childhood and youth (Cohen et al., 1992).

The LNNB-CR was developed on the basis of the neurodevelopmental stages of the child

(Golden, 1981, 1997) and was revised four times in the process (Plaisted, Gustavson, Wilkening, & Golden, 1983). It is designed for children ages 8-12 years and in addition to IQ and achievement provides information specific to motor, rhythm, tactile, visual, receptive speech, expressive language, and memory functions. A description of the LNNB-CR is provided in Table 3. The development of the LNNB-3 represents an extensive revision and major expansion of the LNNB-CR and the adult version. It includes tasks from the previous two measures, but also includes additional tasks, with a total of 27 domains being evaluated. With this major revision, both lower level and more complex tasks and items have been added. The LNNB-3 is intended for use with individuals from age five through adulthood. Interpretation of the LNNB-CR and LNNB-3 focuses predominantly on scale patterns and intrascale (intraindividual) differences, as opposed to levels of performance or pathognomonic signs.

Due to its recent development, there is little research available on the LNNB-3 and most is specific to adults (e.g., Crum, Bradley, Teichner, & Golden, 1997; Crum, Golden, Bradley, & Teichner, 1997). Extensive research has, however been completed with the LNNB-CR. Factor analytic studies (e.g., Karras, Newton, Franzen, & Golden, 1987; Pfeiffer, Naglieri, & Tingstrom, 1987; Sweet, Carr, Rossini, & Kasper, 1986) have resulted in varying factor structures. It has been determined consistently that the LNNB-CR offers unique information not otherwise obtained in psychoeducational assessment, with particular sensitivity to deficits in language, writing, reading, and rhythm (Geary & Gilger, 1984). The pathognomonic scale of the LNNB-CR has been found to account for increased variance, independently of the WISC-R, and to be a better predictor of academic achievement in spelling and reading (McBurnett, Hynd, Lahey, & Town, 1988). It was also found that the LNNB-CR had greater shared variance than the WISC-R with measures of achievement (Hale & Foltz, 1982). The LNNB-CR has been found to be more sensitive to improvement in functioning following medical intervention (e.g., shunt placement) than either cognitive or achievement measures (Torkelson, Liebrook, Gustavson, & Sundell, 1985), as well as supporting differential diagnosis (Carr, Sweet & Rossini, 1986) and the understanding of academic deficits in children with emotional/behavioral problems (Tramontana, Hooper, Curley, & Nardolillo, 1990). The utility of the LNNB-CR in the differentiation of learning disability as opposed to other forms of brain damage, however, has been questioned (e.g., Morgan & Brown, 1988; Oehler-Stinnett,

Table 1 Halstead-Reitan Neuropsychological Battery for Children (ages 9–14 years).

<i>Subtest</i>	<i>Description</i>	<i>Function(s) assessed</i>
Category test	Requires individual to select colors or numbers corresponding to some abstract problem-solving criteria. Immediate feedback is provided for both correct and incorrect responses.	This task assesses general abstraction and concept formation as well as general neuropsychological functioning (Reitan & Wolfson, 1985, 1988).
Tactual performance test	The individual is blindfolded and required to place blocks in slots on a form board using the dominant hand, the nondominant hand, and both hands together. The individual is then asked to draw a diagram of the board with the blocks in their proper spaces.	This task measures tactual discrimination, sensory recognition, and spatial memory. The drawing component is a measure of incidental learning/memory (Reitan & Wolfson, 1988; Selz, 1981).
Speech sounds perception	A taperecorded voice presents a sequence of 60 spoken nonsense words from which the individual must select the correct word each time from three written choices.	This task measures alertness, attention/concentration, and verbal ability (Reitan & Wolfson, 1988)
Seashore rhythm test	The individual is required to differentiate between 30 pairs of rhythmic patterns which are sometimes the same and sometimes different.	This test is thought to be an indicator of generalized cerebral function as well as a measure of alertness and attention/concentration. (Reitan & Wolfson, 1988)
Trail making test	This test uses two tracking tasks, one with numbers (A) and one with letters and numbers (B). First, the individual must connect numbered circles in order; then, the individual must connect circles in sequence, alternating numbers and letters.	The test is believed to measure conceptual flexibility, symbolic recognition, and visual tracking under time constraints (Selz, 1981). It is also used as a measure of overall functioning (Reitan & Wolfson, 1985, 1988).
Finger oscillation test	This test requires the individual to depress a lever as quickly as possible with the index finger of each hand.	This measures motor speed and manual dexterity (Selz, 1981) and lateral dominance (Reitan & Wolfson, 1988.)
Aphasia screening test	This test includes enunciation of spoken language (repeating), naming, reading, writing, spelling, and arithmetic. It also includes copying of a square, circle, and Greek cross.	It is a measure of verbal ability. The drawings are indicative of the verbal-to-motor process (Reitan & Wolfson, 1988).
Sensory perceptual examination		All measure receptive sensory function (Reitan & Wolfson, 1985, 1988).
Tactile perception	The individual is asked to report whether right hand, left hand, right side of face, or left side of face is touched; touches are done unilaterally and bilaterally.	
Auditory perception	Examiner lightly rubs fingers together at the individual's right, left or both ears and the individual is asked to localize the sound produced.	
Visual perception	The individual is asked to report peripheral, unilateral and bilateral single movements produced by the examiner, to assess all four quadrants of the visual field.	

Table 1 (continued)

<i>Subtest</i>	<i>Description</i>	<i>Function(s) assessed</i>
Tactile form recognition	The individual must identify a cross, triangle, square or circle when put in the dominant hand behind a board (unseen) and point to that same object with the nondominant hand; the same process is then carried out with the object in the nondominant hand.	This test is believed to measure tactile perception as well as attention (Nussbaum & Bigler, 1997).
Fingertip number writing	This requires the individual to identify numerals written on their fingertips (both hands).	This is a measure of sensory perceptual functioning (Reitan & Wolfson, 1988).
Grip strength test	Using a hand dynamometer, the strength of grip for the dominant and nondominant hand is determined.	This measures motor functioning and lateral dominance (Reitan & Wolfson, 1988)

Stinnett, Wesley, & Anderson, 1988; Snow & Hynd, 1985b; Snow, Hynd, & Hartlage, 1984). Hynd (1992) also questioned the appropriateness of the standardization sample.

More recently, the Neuropsychological Investigation for Children (NEPSY; Korkman, Kirk, & Kemp, 1997) has been developed for young children. Based on Luria's model (1970), the NEPSY consists of 27 subtests that are summarized in test profiles of strengths and weaknesses. Initially developed in Finnish, in its English version the test includes subtests specific to attention and executive functions, and language, sensorimotor, visuospatial and memory/learning functions (Korkman et al., 1997). It is intended for use with children age 3–12 years. Early research on the NEPSY appears positive (e.g., Korkman, 1988; Korkman, & Hakkinen-Rihu, 1994; Korkman, Liikanen, & Fellman, 1996); however, additional research with this measure, particularly in comparison to the KABC for younger children, will be needed following publication.

In addition, a new battery, the Dean–Woodcock Neuropsychological Assessment System (DWNAS; Dean & Woodcock, in press) is in the process of development. Based on the work of Catell and Horn (Horn, 1988, 1991), this battery combines the cognitive battery of the Woodcock–Johnson Psychoeducational Battery-Revised with a newly developed battery of sensorimotor tests, a structured interview, and a mental status exam. The sensory and motor portion is projected to include eight tests of sensory function and nine tests of motor function. Although some of these tests are similar to those on other neuropsychological batteries, standardized administration, objective scoring criteria, and normative data will be

provided for all tasks. The structured interview and mental status exam are intended to provide information specific to emotional state, motivation, temperament, and prior medical conditions as well as to premorbid history, age at onset and emotional reaction (coping) that may influence neuropsychological performance (Dean & Woodcock, in press). It is projected that this battery will be available in both English and Spanish, with general as well as focused norms that account for age and education using regression methods. As with the NEPSY, clinical research with the DWNAS will be needed once all components of the system are available.

Nomothetic approaches may also be eclectic and use selected measures to sample behaviors from the differing functional systems. Several examples of eclectic batteries can be found in the research literature (e.g., Nussbaum et al., 1988; Rourke, 1994). Since the combinations of measures vary in an eclectic battery from clinical setting to clinical setting, research regarding the efficacy of any given combination of tasks as compared to other combinations or to the published batteries is not feasible, and factor analytic studies of eclectic batteries are not routinely found in the research literature.

4.10.2.5.2 Idiographic approaches

At the other end of the continuum, an idiographic approach tailors the assessment battery to the referral question and the child's individual performance on initial measures administered (Christensen, 1975; Luria, 1973). This type of approach is intended to isolate neurobehavioral mechanisms that underlie the problem of a particular individual rather than

Table 2 Reitan Indiana Neuropsychological Battery (ages 5–8).

<i>Subtest</i>	<i>Description</i>	<i>Function(s) assessed</i>
Category Test	Requires the individual to select colors or numbers corresponding to some abstract problem-solving criteria. Immediate feedback is provided. This version has fewer items and only five categories.	This task assesses general abstraction and concept formation as well as general neuropsychological functioning. (Reitan & Wolfson, 1985).
Matching pictures test	The child matches a single picture to the same picture or to a picture from a more general category.	This task assesses abstraction and concept formation (Reitan & Wolfson, 1988).
Color form test	The child must alternately touch shapes and colors.	This task assesses abstraction and concept formation (Reitan & Wolfson, 1988).
Progressive figures test	The child must use the small shape within a large shape to select the outer shape of the next figure in sequence in a timed condition.	This task assesses concept formation (Reitan & Wolfson, 1988). Also involves cognitive flexibility and attention (Nussbaum & Bigler, 1997)
Tactual performance test	The individual is blindfolded and required to place blocks in slots on a form board using the dominant hand, the nondominant hand, and both hands together. The individual is then asked to draw a diagram of the board with the blocks in their proper spaces.	This task measures tactual discrimination, sensory recognition, and spatial memory. The drawing component is a measure of incidental learning (Reitan & Wolfson, 1988; Selz, 1981).
Finger oscillation test	This test requires the individual to depress a lever as quickly as possible with the index finger of each hand.	This measures motor speed and manual dexterity (Selz, 1981) and lateral dominance (Reitan & Wolfson, 1988.)
Fingertip symbol writing	This requires the individual to identify Xs and Os written on their fingertips (both hands).	This is a measure of sensory perceptual functioning and attention (Reitan & Wolfson, 1988).
Marching test	The child is required to touch a sequence of circles as quickly as possible.	This measures motor functioning (Reitan & Wolfson, 1988).
Sensory perceptual examination		All measures are sensitive to receptive sensory function (Reitan & Wolfson, 1985, 1988).
Tactile perception	The individual is asked to report whether right hand, left hand, right side of face, or left side of face is touched; touches are done unilaterally and bilaterally.	
Auditory perception	Examiner lightly rubs fingers together at the individual's right, left or both ears and the individual is asked to localize the sound produced.	
Visual perception	The individual is asked to report peripheral, unilateral and bilateral single movements produced by the examiner, to assess all four quadrants of the visual field.	
Grip strength test	Using a hand dynamometer, the strength of grip for the dominant and nondominant hand is determined.	This measures motor functioning and lateral dominance of the upper body (Reitan & Wolfson, 1988).

Table 2 (continued)

<i>Subtest</i>	<i>Description</i>	<i>Function(s) assessed</i>
Tactile form recognition	The individual must identify a cross, triangle, square, or circle when put in the dominant hand behind a board (unseen) and point to that same object with the nondominant hand; the same process is then carried out with the object in the nondominant hand.	This test is believed to measure tactile perception as well as attention (Nussbaum & Bigler, 1997).
Aphasia screening test	This test includes enunciation of spoken language, naming, reading, writing, and arithmetic; naming, identifying body parts, left/right, numerals and letters. It also includes drawing a square circle, and Greek cross. It is an abbreviated version of the screening test for older children and adults.	It is a measure of verbal ability. The drawings are indicative of the verbal-to-motor process (Reitan & Wolfson, 1988).
Individual performance Matching Vs, figures Star, concentric squares	Child must match figures or Vs. Child must copy figures of varying difficulty.	These measure visual perceptual and spatial abilities (Reitan & Wolfson, 1988).

providing a detailed evaluation of all areas of functioning. With no predetermined uniformity across evaluations and the dependence on the individual's presenting problems, this approach requires substantial clinical knowledge to determine the components of the battery in order to meet this goal. Due to the limited data on neuropsychological functions and organization of behavior in children, this approach is less frequently used (Fennell, 1994). However, it may be more cost-effective because of the small number of domains which are assessed (Goldstein, 1997). A major drawback to the idiographic approach is the limited research base which is generated and the inability to verify or study the efficacy of this approach as compared to other approaches.

4.10.2.5.3 Combined approaches

The most frequently used approach represents a combination of the nomothetic and idiographic approaches and has been referred to as the flexible battery approach (Sweet et al., 1996). A core set of the same tests is administered to all children, as in the nomothetic approach, and serves as the basis for initial hypothesis generation; this may constitute an initial screening battery. To this core set, further tests are added that are specific to the referral question or that are believed, based on initial observations and performance, to enhance the information provided (Bauer, 1994; L. C. Hartlage & Telzrow,

1986; Rourke, Fisk, & Strang, 1986). The components of the flexible battery itself generally reflect the theoretical position taken by the neuropsychologist with regard to the manner in which behavioral performance reflects brain pathology and the reasons for referral for neuropsychological evaluation for a given individual (Bauer, 1994). According to surveys completed in the 10 years since 1987, the flexible battery approach is generally that preferred by neuropsychologists working with populations of varying ages (Sweet & Moberg, 1990; Sweet et al., 1996). This approach is believed to more accurately identify specific deficits (Batchelor, 1996b).

The Boston Process Approach (Kaplan, 1988; Lezak, 1995) is one example that incorporates a flexible battery. Specific measures with low specificity are used to assess multiple constructs in a variety of neuropsychological domains (Batchelor, 1996b). Hypotheses are then made based on the initial measures, and additional measures with higher levels of specificity are then selected and used to differentiate within and between various functions. Hypotheses initially generated from the screening battery are thus either confirmed or nullified. Inferences are then made regarding brain function based on the specific deficits identified. The flexible battery used in the Boston Process Approach is not limited to quantitative data but also includes qualitative information that is believed to be important in

Table 3 Luria Nebraska Neuropsychological Battery-Children's Revision.

<i>Scale</i>	<i>Description</i>	<i>Function(s) assessed</i>
C1 (motor)	Items cover a variety of motor skills (bilateral and unilateral) including simple hand movements, drawing, and constructional skills.	These tests measure motor domains but are sensitive to many types of motor problems. (Golden, 1997).
C2 (rhythm)	Items include a variety of tasks in which the child is required to report whether one of two groups of tones is higher or lower, reproduce tones and rhythmic patterns, and identify the number of beeps in groups of sounds.	These items are considered to be most sensitive to attention and concentration (Golden, 1997).
C3 (tactile)	Items include tasks in which the child is asked to report where they are touched, how hard they are touched, as well as to name and identify objects through touch.	These items measure the extent of cutaneous sensation and stereognostic perception (Golden, 1997).
C4 (visual)	Items include tasks in which the child is required to identify an object or picture, overlapping pictures, pictures that are difficult to perceive, and mirror image versions; items also include progressive matrices, and spatial rotation.	These items measure visual-spatial organization and perception as well as right hemisphere function (Golden, 1997).
C5 (receptive speech)	The child is required to repeat phonemes, repeat phonemes at various levels of pitch, name objects, point to objects, identify and define words, and respond to sentences.	These items measure receptive language and auditory skills as well as left hemisphere function (Golden, 1997).
C6 (expressive speech)	The child is required to repeat phonemes, words, and sentences as well as to generate speech forms including naming objects, counting forward and backward, spontaneous discourse in response to a picture, story, or discussion topic.	These items measure expressive language as well as left hemisphere function. Results may be impacted by reading ability (Golden, 1997).
C7 (writing)	Tasks include copying of letters and words, writing first and last name, writing sounds, words, and phrases from dictation.	These items measure visual motor and auditory motor skills and are believed to measure functioning of the temporal-parietal-occipital area (Golden, 1997).
C8 (reading)	The child is asked to generate sounds from letters, name letters, read simple words, sentences, and paragraphs.	These items measure reading as well as left hemisphere function (Golden, 1997).
C9 (arithmetic)	Child is asked to write arabic and roman numerals from dictation, compare numbers, complete simple computation problems, and generate serial threes.	These tasks measure arithmetic skills, but are considered the most sensitive to educational deficits as well to all/any dysfunction (Golden, 1997).
C10 (memory)	Tasks required include having the child memorize words as well as predicting their own performance on various memory tasks.	These items measure short-term memory functions and are most sensitive to verbal dysfunction (Golden, 1997).
C11 (intellectual)	The child is asked to complete a variety of tasks including interpretation of pictures, arranging pictures in order, identification of what is comical/absurd, interpretation of story, determination of similarities, simple arithmetic problems, identification of logical relations and so on.	These are considered to be reflective of general neuropsychological function, concept formation, and reasoning (Golden, 1997).

understanding the child's problems and in developing effective intervention programs (Batchelor, 1996b; Milberg, Hebben, & Kaplan, 1996). There is less of a focus on the results of standardized test performance with greater attention paid to developmental history, presentation of symptoms, strategy use in task completion, and error analysis. As such, the "process" approach uses both standardized measures and experimental measures as well as "testing of limits" that may involve procedural modifications in order to gain insight into brain-behavior relationships (Kaplan, 1988; Milberg, et al., 1996). Concern has, however, been expressed regarding the reliability of scores obtained on standardized measures when the standardization procedures have been compromised (e.g., Rourke et al., 1986). Further, most of the research and clinical study, with the Boston Process Approach in particular, has been with adult populations as opposed to children, and it is not recommended for other than research applications.

4.10.2.6 General Organization of the Neuropsychological Assessment of the Child

When the neurologist examines a child, the physical examination looks principally for structural defects in the CNS, trauma to the CNS, or specific disease entities or toxins. An assessment of history is an integral component of both the neurological and the neuropsychological assessment of children and includes, for the neurologist, assessment of the gestational period, delivery, postnatal history, and the family medical history through at least two generations. The physical examination that lay people view as the neurological examination proper is based largely on observations of the neurologist and is conducted in the context of a brief interview and physical manipulation to assess tone, muscle strength, deep tendon reflexes, sensation, and brain stem and spinal reflexes. Electrophysiologic, serologic and/or imaging studies may then be ordered as may be suggested by such results. Neuropsychological testing may also be ordered when there are suspicions of intellectual delay or functional sequelae are suspected, related to trauma, disease, or toxins. As recently as the 1970s and into the 1980s, neuropsychological testing was used to evaluate lesion site and size and to assist in the differential diagnosis of a variety of neurologic diseases, but this function has been largely supplanted by advances in neuroimaging, clinical serology, and the linking of a variety of cancers to mental symptoms.

The neuropsychological examination of children is focused more directly on an analysis of the functional concomitants and sequelae than on the identification of strictly neurologic disorders, but it is also useful in the diagnosis and identification of more subtle conditions (e.g., learning disabilities, ADHD) or other neurologic disorders, especially in early stages (such as childhood onset of Huntington's disease) that are more resistant to diagnosis via neurologic examination. (In adulthood, differential diagnosis, such as depression versus dementia or differentiation of malingering or among various dementias takes on greater importance.) At all ages, the neuropsychological examination is also focused on rehabilitation.

A thorough history is important to a proper neuropsychological assessment. The length of time since trauma or disease onset, premorbid levels of functions, family history of related problems, and problems related to gestation, delivery, and the postnatal period are all relevant to accurate interpretation of the results of neuropsychological testing. If a school-aged or college-aged individual is involved, it is important to review educational history with specific performance data including standardized tests scores and grades along with any special education history. With children who are, developmentally, a moving target, it is important to be always cognizant of the educational implications of the reason for referral. Following a review of history and obtainable records, there are nine key points to consider in the organization of the neuropsychological assessment.

(i) *All or at least a significant majority of the child's educationally relevant cognitive skills or higher order information process skills should be assessed.* This will often involve an assessment of general intellectual level via a comprehensive IQ test such as a Wechsler scale or KABC. Evaluation of the efficiency of mental processing as assessed by strong measures of *g*, is essential to provide a baseline for interpreting all other aspects of the assessment process. Assessment of basic academic skills including reading, writing, spelling, and math will be necessary, along with tests of memory and learning such as the TOMAL which also have the advantage of including performance-based measures of attention and concentration. Problems with memory, attention and concentration, and new learning are the most common of all complaints following CNS compromise and are frequently associated with more chronic neurodevelopmental disorders (e.g., learning disability, ADHD).

(ii) *Testing should sample the relative efficiency of the right and of the left hemispheres of the brain.* Asymmetries of performance are of

interest in their own right, but different brain systems are involved in each hemisphere that have different implications for treatment as well. Even in a diffuse injury such as anoxia, it is possible to find greater impairment in one portion of an individual's brain than in another. Specific neuropsychological tests like those of Halstead and Reitan or the LNNB-CR are useful here.

(iii) *Sample anterior and posterior regions of cortical function.* The anterior portion of the brain is generative and regulatory while the posterior region is principally receptive. Deficits and their nature in these systems will have great impact on treatment choices. Many common tests such as receptive (posterior) and expressive (anterior) vocabulary tests may be applied here along with a systematic and thorough sensory perceptual examination. In conjunction with key point (ii), this allows for evaluation of the integrity of the four major quadrants of the neocortex.

(iv) *Determine the presence of specific deficits.* Any specific functional problems a child is experiencing must be determined and assessed. In addition to those being of importance in the assessment of children with neurodevelopmental disorders, traumatic brain injury (TBI), stroke, even some toxins can produce very specific changes in neocortical function that are addressed best by the neuropsychological assessment. Similarly, research with children with leukemia suggests the presence of subtle neuropsychological deficits following chemotherapy that may not be detected by more traditional psychological measures. Neuropsychological tests tend to be less g-loaded as a group and to have greater specificity of measurement than many common psychological tests. Noting areas of specific deficit is important in both diagnosis and treatment planning.

(v) *Determine the acuteness versus the chronicity of any problems or weaknesses found.* The "age" of a problem is important to diagnosis and to treatment planning. Combining a thorough history with the pattern of test results obtained, it is possible, with reasonable accuracy, to distinguish chronic neurodevelopmental disorders such as dyslexia or ADHD from new acute problems resulting from trauma, stroke, or disease. Care must be taken especially in developing a thorough, documented history when such a determination is made. When designing intervention/treatment strategies, rehabilitation and habilitation approaches take differing routes depending upon the age of the child involved and the acuteness or chronicity of the problems evidenced.

(vi) *Locate intact complex functional systems.* It is imperative in the assessment process to

locate strengths of the child and intact systems that can be used to overcome the problems the child is experiencing. Treatment following CNS compromise involves habilitation and rehabilitation with the understanding that some organic deficits will represent permanently impaired systems. As the brain is a complex interdependent systemic network of complex organizations that produce behavior, the ability to identify intact systems is crucial to enhancing the probability of designing successful treatment. Identification of intact systems also suggests the potential for a positive outcome to parents and teachers, as opposed to fostering low expectations and fatalistic tendencies on identification of brain damage or dysfunction.

(vii) *Assess affect, personality, and behavior.* Neuropsychologists sometimes ignore their roots in psychology and focus on assessing the neural substrates of a problem. However, CNS compromise will result in changes in affect, personality, and behavior. Some of these changes will be transient, some will be permanent, and due to the developmental nature of children, some will be dynamic. Some of these changes will be direct (i.e., a result of the CNS compromise at the cellular and systemic levels) and others will be indirect (i.e., reactive to loss or changes in function, or to how others respond to and interact with the individual). A thorough history, including onset of problem behaviors, can assist in determination of direct versus indirect effects. Comprehensive approaches such as the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992) which contain behavior rating scales, omnibus personality inventories, and direct observation scales seem particularly useful. Such behavioral changes will also require intervention and the latter may vary depending on whether the changes noted are direct or indirect effects or whether there were behavior problems evident on a premorbid basis.

(viii) *Test results should be presented in ways that are useful in school settings, not just in acute care or intensive rehabilitation facilities.* Schools are a major context in which children with chronic neurodevelopmental disorders must function. Children who have sustained insult to the CNS (i.e., TBI, stroke) will eventually return to a school or similar educational setting. Schools are where the greatest long-term impact on a child's outcome after CNS compromise is seen and felt. Results should speak to academic and behavioral concerns, reflecting what a child needs to be taught next in school, how to teach to the child's strengths through the engagement of intact complex functional systems, and how to motivate and manage positive behavioral outcomes. For children with TBI, additional

information regarding potential for recovery and the tenuousness of evaluation results immediately post-injury needs to be communicated as does the need for reassessment of both the child and the intervention program at regular intervals.

(ix) *If consulting directly to a school, be certain the testing and examination procedures are efficient.* School systems, which is where one finds children, do not often have the resources for funding the type of diagnostic workups neuropsychologists prefer. Therefore, when consulting to the school, it is necessary to be succinct and efficient in planning the neuropsychological evaluation. If the school can provide the results of a very recent intellectual and academic assessment as well as the behavioral assessment information, this can be then integrated into the neuropsychological assessment by the neuropsychologist. If a recent intellectual and academic assessment has not been completed, it may be cost-efficient for qualified school district personnel to complete this portion of the assessment for later integration with other data obtained and interpreted by the neuropsychologist. For children in intensive rehabilitation facilities or medical settings, it may be appropriate for school personnel to participate in the evaluation prior to discharge (i.e., for children with TBI being released and returned to the schools). This collaborative involvement can facilitate program planning with the receiving school district and is preferable to eliminating needed components of the neuropsychological evaluation.

When considering rehabilitation of the child with a focal injury or TBI, several additional considerations are evident. It is important to determine what type of functional system is impaired. Impaired systems may, for example, be modality-specific or process-specific. The nature or characteristics of the impairments must be elucidated before an intelligent remedial plan can be devised.

The number of systems impaired should be determined and prioritized. Children may not be able to work out everything at once and a system of priorities should be devised so that the most important of the impairments to impact overall recovery is the first and most intensely addressed area of impairment. The degree of impairment, a normative question, is also an important consideration in this regard. At times, this will require the neuropsychologist to reflect also on the indirect effects of a TBI, as an impaired or dysfunctional system may adversely affect other systems that are without true direct organic compromise.

The quality of neuropsychological strengths that exist will also be important and tends to be

more of an ipsative as opposed to a normative determination. Furthermore, certain strengths are more useful than others. Preserved language and speech are of great importance for example, while an intact sense of smell (an ability often impaired in TBI) is of less importance in designing treatment plans and outcome research. Even more important to long-term recovery are intact planning and concept formation skills. The executive functioning skills of the frontal lobes take on greater and greater importance with age, and strengths in those areas are crucial to long-term planning (as is the detection of weaknesses). These will change, however, with age as the frontal lobes become increasingly prominent in behavioral control after age nine years, again through puberty, and continuing into the 20s.

There are of course times when the scope of the neuropsychological assessment of a child is less broad. On occasion, referrals may be very specific (e.g., "Does Susan have memory or attention problems?"). Even when such seemingly succinct questions are asked, it is commonly a good practice to inquire of the referral source as to whether other questions may be anticipated (e.g., Is memory an issue because of poor school achievement? Possible learning disability?).

This section draws in part upon the writings, teachings, and workshops of Lawrence C. Hartlage and Byron Rourke.

4.10.2.7 Interpretation Issues

Neuropsychological assessment of children yields not only an accumulation of test data and impressions, but also a variety of paradigms for understanding and interpreting that data. There are a number of competing paradigms and theories (e.g., Ayers, 1974; Das et al., and 1979; Luria, 1966; Reynolds, 1981b), and as a result, not only is there considerable variability in the quality and choice of measures used in the neuropsychological assessment of children, there are considerable differences in the ways in which the data obtained are used for making inferences and eventually interpreted (Batchelor, 1996a; Nussbaum & Bigler, 1997). Interpretation of the accumulated data is dependent to a great extent on the neuropsychologist's clinical skills and acumen (D'Amato et al., 1997). Interpretation may be based on overall performance level (e.g., Reitan, 1986, 1987), performance patterns (e.g., Mattarazzo, 1972; Reitan, 1986, 1987), asymmetry of function (e.g., L. C. Hartlage, 1982), the presence of "organic" signs (Kaplan, 1988; Lezak, 1995), or on some combination of features. It is not necessarily

the case that only one paradigm is appropriate; which paradigm is most suitable may depend on the child being evaluated. Most importantly, the model used for interpretation should allow the neuropsychologist to make predictions about the child's ability to perform in a variety of contexts and about the efficacy of treatment/intervention plans (Reynolds et al., 1997).

4.10.2.7.1 Performance level

With the use of this indicator, the child's overall level of performance is compared to normative data and conclusions are reached based on deviations from the norm. The extent of variability among typically developing children on some measures at given ages (e.g., when the standard deviation approximates the mean score) may preclude interpretation of results using this approach. In addition, this approach can be misleading, particularly in those individuals with higher cognitive ability (Jarvis & Barth, 1984; Reitan & Wolfson, 1985). Further, there is a tendency for this method to yield a large number of false positives due to the potential for other factors (e.g., motivation, fatigue) to impact on a child's performance (Nussbaum & Bigler, 1997).

4.10.2.7.2 Profile patterns

Application of the neuropsychological model to learning problems has been criticized as being too aligned with a medical model and an emphasis on pathology (Gaddes & Edgell, 1994). As asserted by Little and Stavrou (1993), merely identifying that brain integrity has in some way been compromised is not in and of itself particularly helpful to the child or to those who need to develop interventions to help the child. Neuropsychologists look beyond diagnosis or categorization to an understanding of brain-behavior relations. In order to accomplish this, neuropsychological assessment involves consideration of associations and dissociations of performance across measures (Fletcher, 1988; Rutter, 1981). Performance patterns or intraindividual differences provide a means of conceptualizing functional vs. dysfunctional organizational systems. Strengths and weaknesses are then identified based on the discrepancies between the domains studied. This method has, however, been used frequently for the identification or classification of subtypes of learning disabilities (Branch et al., 1995; Nussbaum & Bigler, 1986; Rourke, 1984).

In interpreting data obtained using this type of evaluation, clinicians differ with regard to emphasis on child strengths, child weaknesses or a combination of strengths and weaknesses

(L. C. Hartlage & Telzrow, 1983; Reynolds, 1981b, 1986a; Teeter, 1997). It has been argued that a strength model is more efficacious, with habilitation based on those complex functional systems that are sufficiently intact, and therefore potentially capable of taking over and moderating the acquisition of the skills needed (Reynolds et al., 1997). Emphasis on weaknesses, generally referred to as the deficit model, is not supported by research (e.g., Adams & Victor, 1977; L. C. Hartlage, 1975; P. L. Hartlage & Givens, 1982; P. L. Hartlage & Hartlage, 1978), and deficit approaches to intervention (e.g., remediation of the deficit process) have not been found to be effective and may even be harmful (L. C. Hartlage & Reynolds, 1981).

There are, however, some problems with this method of interpretation regardless of whether the focus is on strengths, weaknesses, or a combination of these. This approach may be misleading as other variables may account for these intra-individual differences (Jarvis & Barth, 1984). Additionally, some such intra-individual differences (e.g., verbal IQ-performance IQ differences) have been found to occur with frequency in the general population (e.g., Kaufman, 1976b) and seemingly abnormal levels of subtest scatter (WISC-R) have been found to be relatively common (Gutkin & Reynolds, 1980; Kaufman, 1976a, 1976b; Reynolds, 1979). Base rates in the general population of specific intra-individual differences for various other combinations of measures have not been studied, and what appears to be a "difference" may not be unusual or unique at a given age level. Further, the stability of these profile patterns over at least very short periods of time needs to be investigated (Reynolds, 1997b).

4.10.2.7.3 Functional asymmetry

Examination of asymmetries in performance across measures is another method of intra-individual consideration. Replicable asymmetries in performance are generally considered signs of CNS dysfunction (Batchelor, 1996b). Most frequently, the comparison is made between those functions that are believed to be right hemisphere-dominated as opposed to left hemisphere-dominated. These differences, however, may be difficult to interpret, particularly for younger children (Reynolds et al., 1997). Further, understanding of the lateralization of cortical functions is frequently based on evidence from adults as opposed to children and assumes that the lateralization is stable over time, despite differing rates of brain maturity (Spreeen et al., 1995). Reliance on left-right

differences and measures based on lateralization of function have also been criticized as ignoring the role of hemispheric interaction on behavior (e.g., Efron, 1990; Hiscock & Kinsbourne, 1987).

As with the patterns of performance method, right-left differences have been used in the characterization of children with right hemisphere dysfunction as suggestive of learning disability or ADHD (Rourke, 1989). The results of studies are, however, equivocal (e.g., Branch et al., 1995; Gross-Tsur, Salev, Manor, & Amir, 1995; Voeller, 1995). Research, in general, regarding lateralization of function and hemispheric specialization is fraught with conflicting results (e.g., Bever, 1975; Das et al., 1979; Dean, 1984; Reynolds, 1981b), and it has been suggested that the traditional verbal-nonverbal distinction between hemispheres is an oversimplification of a complex system (Dean, 1984; Reynolds, 1981a). Based on Luria's theories, asymmetries of function are not content- or modality-specific but rather are "process"-specific (Reynolds, 1981a, 1981b). Bever (1975) posited two fundamental lateralized processing types, the analytic and holistic; these were translated into sequential and simultaneous in the KABC based on Das et al. (1979). In the research literature, however, there is often a preponderance of emphasis placed on content and modality, as opposed to process, in the interpretation of functional asymmetries; it is believed that this may account for the conflicting results across studies (Reynolds et al., 1997).

Nussbaum et al., (1988) proposed an alternate method of examining asymmetry. In the model of Nussbaum and colleagues, the neuropsychological protocol and interpretation reconceptualized neurobehavioral functioning along the anterior-posterior gradient as opposed to left-right differences. The recommended protocol includes tasks from the HRNB as well as from other batteries. Nussbaum and colleagues asserted that this model may provide additional information in the investigation of asymmetries in children with learning and behavioral problems. Initial research in this area suggested that weaknesses on anterior measures were associated with psychological/behavioral problems (Teeter & Semrud-Clikeman, 1997). Two later studies, however, failed to support the anterior-posterior gradient theory (Mataszow & Hynd, 1992).

4.10.2.7.4 Pathognomonic signs

The pathognomonic signs approach involves the identification of specific deficits or performance errors that are not frequently found in typically developing individuals (Nussbaum &

Bigler, 1997). Identification of "organic" signs is generally completed through qualitative analysis of errors (Kaplan, 1988; Lezak, 1995). The presence of specific types of errors is then seen as an indication of a compromise to brain integrity. This method has been used reliably with adult populations; however, the utility of this approach in the neuropsychological assessment of children has not been demonstrated (Batchelor, 1996b). The range of variability associated with the developmental process in children would seem to make it more difficult to interpret specific errors as signs of organic impairment (Nussbaum & Bigler, 1997). Unlike the performance levels method, the use of pathognomonic signs has been found to result in a large number of false negatives (Boll, 1974). This may be related to the potential for reorganization/recovery of function in children (Nussbaum & Bigler, 1997).

4.10.2.7.5 Combination approaches

Boll (1981) proposed utilizing performance levels, patterns of performance, pathognomonic signs, and asymmetry of function in concert, in order to account for the potential limitations to the use of any single approach in the interpretation of neuropsychological assessment data. This multiple inferential levels approach is used in the HRNB and is supported by others as well (e.g., Rourke, 1994). The "rules approach" (Selz & Reitan, 1979b) also combines approaches, but in a different manner. Using the "rules approach," each of 37 aspects of neuropsychological performance is rated on a four-point scale in order to provide an objective system for measuring the extent of impairment.

More recently, Taylor and Fletcher (1990) proposed that the child's performance on neuropsychological measures be used to identify and clarify the functional aspects of the child's problems, with the understanding that the biological or neurological substrates of the learning or behavior problem serve to set limits on the child's performance. Levine (1993) has posited still another model for interpretation of neuropsychological data. The "observable phenomenon" model places the emphasis on observable behaviors, that may impact on classroom performance and the changing demands placed on the child over time, as opposed to test results.

4.10.3 CONCLUSIONS

Neuropsychological assessment and the field of clinical child neuropsychology in general have much to offer in the way of understanding

the functional systems of the brain and the mechanisms involved in the learning and self-regulation process. Not only is this important in understanding and designing treatment programs for children with problems, but increased understanding of brain functions and their relation to behavior can also improve the overall outcomes for all children (Gaddes, 1983). Historically, neuropsychological assessment of children has taken its lead from research and practice with adults. Issues relating to neurodevelopment, task appropriateness, varying contexts for children, progression following brain injury with children, and so on, render the continuing use of this approach inappropriate. A variety of theoretical models exist; however, many of these are adult-based and used without consideration of developmental issues. Only by developing its own theories and clinical assessment procedures, that are sensitive to developmental features and responsive to educational issues, can the field of clinical child neuropsychology continue to advance and make meaningful contributions to the understanding of learning and behavior problems in children. Development and incorporation of typologies within the childhood disorders, based on clinical experience with children and neuropsychological theory that addresses habilitation, programming, and research needs and that has intuitive appeal to psychologists, educators, and neurologists, would be viewed as a major conceptual contribution to the field of child neuropsychology (Reynolds, 1986b).

Continued methodological and measurement problems in the research that serves as a foundation for the interpretation of neuropsychological data impede progress in the field of clinical child neuropsychology and impact on the accuracy of diagnosis and on the appropriateness of treatment planning. Lack of attention to standard psychometric methods within the field of clinical neuropsychology is all too rampant and poses serious limitations in research in clinical arenas; intuitive appeal, clinical acumen, and perceived utility are not sufficient, but must be combined with sound empirical research (Reynolds, 1986b). While it is anticipated that new measures will have sufficient normative samples for evaluation of associations with demographic variables, and assessment of validity, reliability, sensitivity, and specificity issues, many existing measures continue to have insufficient normative data. All too often, sensitivity is a focus; specificity is also necessary if results are to be useful in treatment/intervention planning. This requires further investigation of contrasting clinical groups. In research with clinical groups, there is a need to consider comorbidity and be as

specific as possible in descriptions of clinical subgroups (e.g., Fletcher, Shaywitz, & Shaywitz, 1994). Regardless of the perspective used in interpretation, the value of that interpretation is only as good as the measures used in the assessment process and their sensitivity and specificity (Batchelor, 1996b) in combination with the skills and knowledge of the user (Golden, 1997). Failure to resolve these measurement and methodology issues has impeded and will continue to impede progress in the field of neuropsychology (Reynolds, 1997b).

The field of clinical child neuropsychology is in part driven by the development and application of standardized diagnostic procedures that are sensitive to higher cognitive process as related to brain function (Reynolds, 1997b). While the development of new measures of memory, attention, information processing, and so on provide alternatives for clinical child neuropsychologists (e.g., Reynolds, & Bigler, 1994), further research is needed with these as well as with other new and existing measures in order to determine their utility as part of a comprehensive neuropsychological battery. The incorporation of computer-based assessment is likely to increase in the next decades with the potential for incorporation of computer simulation, interactive types of tasks, virtual reality and so on, as means of measuring neuropsychological function. Computerized testing may facilitate the interface with electrophysiological and neuroradiological methods and, ultimately, bring about significant advances in the understanding of learning and behavior problems. Technological advances in assessment, however, will require the same types of research regarding psychometric properties, confounding factors, cultural/gender differences, and so on. Even with those existing tests that currently include an option for computerized assessment, there are some indications of differences in results following computer administration as opposed to the more traditional administration. If this is the case, then it may be appropriate for separate normative data to be obtained for each mode of administration. Furthermore, children with substantial CNS compromise will have difficulty manipulating computerized test materials, and careful validity research will be required at a time when publishers and others are looking for ways to reduce costs associated with health care products.

A major concern with regard to the increased emphasis on reducing health care costs is that neuropsychologists will shorten tests and attempt to streamline batteries, and in the process lessen both the quantity of time required and the quality of the assessment provided (Woody, 1997). This not only impacts on clinical practice,

but also on the knowledge generated through future research. Conflict or dissonance among clinicians will not be well tolerated and the future of clinical child neuropsychology will need the support of public policy (Woody, 1997). This means that reasonable agreement on theoretical foundations, training, and procedures will need to be reached (Woody, 1997). Provision of neuropsychological services needs to be predicated on academics, research, and training consistent with both a clinical psychology orientation and specialized training in brain-behavior relationships, in order to ensure sound foundations (Woody, 1997). With regard to neuropsychological assessment of children, there is a definite need on the part of child neuropsychologists to be well grounded in the developmental process, from a cognitive as well as neurological perspective. Children are a moving target and need even more sophisticated assessment devices than do adults. Considerable work remains in this domain alone.

Neuropsychological assessment of children with learning or behavioral/emotional problems is not necessarily essential. It does, however, provide for a comprehensive evaluation of cognitive skills and emotional factors as well as of environmental influences (L. C. Hartlage & Long, 1997). Neuropsychological assessment may be most appropriate for those children who exhibit characteristics of a disorder that includes cognitive deficits (e.g., learning disability) or significant behavioral problems (e.g., behavior disorder) in the absence of neurophysiological evidence of brain damage, for those with a known neurological syndrome or disease as evidenced by neurophysiological methods, and for those children who, because of genetic predisposition or prenatal/perinatal complications, are believed to be at high risk for neurological disorders (Allen, 1989).

4.10.4 REFERENCES

- Adams, R. L. (1985). Review of the Luria Nebraska Neuropsychological Battery. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook*. Lincoln, NE: University of Nebraska.
- Adams, R. D., & Victor, M. (1977). *Principles of neurology*. New York: McGraw-Hill.
- Allen, C. (1989). Why use neuropsychology in the schools? *The Neuro-Transmitter*, 1, 1-2.
- Ardila, A., & Roselli, M. (1994). Development of language, memory, and visuospatial abilities in 5- to 12-year old children using a neuropsychological battery. *Developmental Neuropsychology*, 10, 97-120.
- Ardila, A., Roselli, M., & Puente, T. (1994). *Neuropsychological evaluation of the Spanish speaker*. New York: Plenum.
- Arffa, S., Fitzhugh-Bell, K., & Black, F. W. (1989). Neuropsychological profiles of children with learning disabilities and children with documented brain damage. *Journal of Learning Disabilities*, 22, 635-640.
- Arnold, B. R., Montgomery, G. T., Castenada, I., & Langoria, R. (1994). Acculturation of performance of Hispanics on selected Halstead-Reitan neuropsychological tests. *Assessment*, 1, 239-248.
- Asarnow, R. F., Asamen, J., Granholm, E., & Sherman, T. (1994). Cognitive/neuropsychological studies of children with a schizophrenic disorder. *Schizophrenia Bulletin*, 20, 647-669.
- Asarnow, R. F., Brown, W., & Strandburg, R. (1995). Children with a schizophrenic disorder: Neurobehavioral studies. *European Archives of Psychiatry and Clinical Neuroscience*, 245(2), 70-79.
- Ayers, A. J. (1974). *Sensory integration and learning disorders*. Los Angeles: Western Psychological Services.
- Aylward, G. P., Gioia, G., Verhulst, S. J., & Bell, S. (1995). Factor structure of the Wide Range Assessment of Memory and Learning in a clinical population. *Journal of Psychoeducational Assessment*, 13, 132-142.
- Bakker, D. J. (1984). The brain as a dependent variable. *Journal of Clinical Neuropsychology*, 6, 1-16.
- Bannatyne, A. (1974). Diagnosis: A note on recategorizations of the WISC scale scores. *Journal of Learning Disabilities*, 7, 272-274.
- Barkley, R. A. (1991). The ecological validity of laboratory and analogue assessments of ADHD symptoms. *Journal of Abnormal Child Psychology*, 19, 149-178.
- Barkley, R. A. (1994). The assessment of attention in children. In G. R. Lyon (Ed.), *Frames of reference of the assessment of learning disabilities: New views on measurement issues* (pp. 69-102). Baltimore: Brookes.
- Barkley, R. A., DuPaul, G. J., & McMurray, M. B. (1991). Attention deficit disorder with and without hyperactivity: Clinical response to three dose levels of methylphenidate. *Pediatrics*, 87, 519-531.
- Barkley, R. A., Fischer, M., Newby, R., & Breen, M. (1988). Development of multimethod clinical protocol for assessing stimulant drug responses in ADHD children. *Journal of Clinical Child Psychology*, 20, 163-188.
- Batchelor, E. S. (1996a). Introduction. In E. S. Batchelor, Jr. & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation* (pp. 1-8). Boston: Allyn & Bacon.
- Batchelor, E. S. (1996b). Neuropsychological assessment of children. In E. S. Batchelor, Jr. & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation* (pp. 9-26). Boston: Allyn & Bacon.
- Batchelor, E. S., Gray, J. W., Dean, R. S., & Lowery, R. (1988). Interactive effects of socioeconomic factors and perinatal complications. *NASA Program Abstracts*, F-169, 115-116.
- Batchelor, E. S., Kixmiller, J. S., & Dean, R. S. (1990). Neuropsychological aspects of reading and spelling performance in children with learning disabilities. *Developmental Neuropsychology*, 6, 183-192.
- Batchelor, E. S., Sowles, G., Dean, R. S., & Fischer, W. (1991). Construct validity of the Halstead Reitan Neuropsychological Battery for children with learning disabilities. *Journal of Psychoeducational Assessment*, 9, 16-31.
- Bauer, R. M. (1994). The flexible battery approach to neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (pp. 259-290). Hillsdale, NJ: Erlbaum.
- Becker, M. G., Isaac, W., & Hynd, G. W. (1987). Neuropsychological development of nonverbal behaviors attributed to "frontal lobe" functioning. *Developmental Neuropsychology*, 3, 275-298.
- Bellinger, D. (1995). Lead and neuropsychological function in children: Progress and problems in establishing brain-behavior relationships. In M. G. Tramontana & S. R. Hooper (Eds.), *Advances in Child Neuropsychology*

- (Vol. 3, pp. 12–47). New York: Springer-Verlag.
- Benton, A. L., Hamsher, K., Varney, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment*. New York: Oxford University Press.
- Bever, T. G. (1975). Cerebral asymmetries in humans are due to the differentiation of two incompatible processes: Holistic and analytic. In D. Aaronson & R. Reiber (Eds.), *Developmental neurolinguistics and communication disorders*. New York: New York Academy of Sciences.
- Beverley, D. W., Smith, I. S., Beesley, P., Jones, J., & Rhodes, N. (1990). Relationship of cranial ultrasonography, visual and auditory evoked responses with neurodevelopmental outcome. *Developmental Medicine and Child Neurology*, 32, 210–222.
- Bigler, E. R. (1990). *Traumatic brain injury: Mechanisms of damage, assessment, intervention and outcome*. Austin, TX: Pro-Ed.
- Bigler, E. R. (1991). Neuropsychological assessment, neuroimaging, and clinical neuropsychology. *Archives of Clinical Neuropsychology*, 6, 113–132.
- Bigler, E. R. (1996). Bridging the gap between psychology and neurology: Future trends in pediatric neuropsychology. In E. S. Batchelor, Jr. & R. S. Dean (Eds.), *Pediatric Neuropsychology* (pp. 27–54). Boston: Allyn & Bacon.
- Black, F. W. (1976). Cognitive, academic, and behavioral findings in children with suspected and documented neurological dysfunction. *Journal of Learning Disabilities*, 9, 182–187.
- Boll, T. J. (1974). Behavioral correlates of cerebral damage in children age 9–14. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and applications* (pp. 91–120). Washington, DC: Winston.
- Boll, T. J. (1978). Diagnosing brain impairment. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders*. New York: Plenum.
- Boll, T. J. (1981). The Halstead Reitan Neuropsychological Battery. In S. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 577–607). New York: Wiley-Interscience.
- Bolter, J. F., & Long, C. J. (1985). Methodological issues in research in developmental neuropsychology. In L. C. Hartlage & C. F. Telzow (Eds.), *Neuropsychology of individual differences: A developmental perspective* (pp. 41–59). New York: Plenum.
- Branch, W. B., Cohen, M. J., & Hynd, G. W. (1995). Academic achievement and attention-deficit/hyperactivity disorder in children with left- or right-hemisphere dysfunction. *Journal of Learning Disabilities*, 28, 35–43.
- Breslau, N., Chilcoat, H., DelDotto, J., & Andreski, P. (1996). Low birth weight and neurocognitive states at six years of age. *Biological Psychiatry*, 40, 389–397.
- Breslau, N., & Marshall, I. A. (1985). Psychological disturbance in children with physical disabilities: Continuity and change in a 5-year follow-up. *Journal of Abnormal Child Psychology*, 13, 199–216.
- Burin, D. I., Prieto, G., & Delgado, A. (1995). Solution strategies and spatial visualization strategies: Design of a computerized test for their assessment. *Interdisciplinaria*, 12(2), 123–137.
- Carr, M. A., Sweet, J. J., & Rossini, E. (1986). Diagnostic validity of the Luria Nebraska Neuropsychological Battery-Children's Revision. *Journal of Consulting and Clinical Psychology*, 54, 354–358.
- Chelune, G. J., & Baer, R. A. (1986). Developmental norms for the Wisconsin Card Sorting Test. *Journal of Clinical and Experimental Neuropsychology*, 8, 219–228.
- Chelune, G. J., & Thompson, L. L. (1987). Evaluation of the general sensitivity of the Wisconsin Card Sorting Test among younger and older children. *Developmental Neuropsychology*, 3, 81–89.
- Christensen, A. L. (1975). *Luria's neuropsychological investigation*. New York: Spectrum.
- Cicchetti, D. V. (1994). Multiple comparison methods: Establishing guidelines for their valid application in neuropsychological research. *Journal of Clinical and Experimental Neuropsychology*, 16, 155–161.
- Clark, E., & Hostettler, C. (1995). *Traumatic brain injury: Training manual for school personnel*. Longmont, CO: Sopris West.
- Cohen, M. J. (1997). *The Children's Memory Scale*. San Antonio, TX: Psychological Corporation.
- Cohen, M. J., Branch, W. B., Willis, W. G., Weyandt, L. L., & Hynd, G. W. (1992). Childhood. In A. E. Puente & R. J. McCaffrey (Eds.), *Handbook of neuropsychological assessment* (pp. 49–79). New York: Plenum.
- Cohen, S. E., Beckwith, L., Parmalee, A. H., & Sigman, M. (1996). Prediction of low and normal school achievement in early adolescents born preterm. *Journal of Early Adolescence*, 16, 46–70.
- Conners, C. (1995). *Continuous performance test computer program 3.0: User's manual*. Toronto, London: Multi-Health Systems.
- Copeland, D. R., Dowell, R. E., Jr., Fletcher, J. M., Bordeaux, J. D., Sullivan, M. P., Jaffe, N., Frankel, L. S., Ried, H. L., & Cangir, A. (1988). Neuropsychological effects of childhood cancer treatment. *Journal of Child Neurology*, 3, 53–62.
- Crockett, D., Klonoff, H., & Bjerring, J. (1969). Factor analysis of neuropsychological tests. *Perceptual and Motor Skills*, 29, 791–802.
- Crum, T. A., Bradley, J. D., Teichner, G., & Golden, C. J. (1997, November). *Analysis of the general intelligence subtest of the Luria Nebraska neuropsychological battery III*. Paper presented at the 17th Annual Conference of the National Academy of Neuropsychologists, Las Vegas, NV.
- Crum, T. A., Golden, C. J., Bradley, J. D., & Teichner, G. (1997, November). *Analyzing the concurrent validity of the memory scales of the Luria Nebraska neuropsychological battery-Third edition*. Paper presented at the 17th Annual Conference of the National Academy of Neuropsychologists, Las Vegas, NV.
- Damasio, A. R., & Maurer, R. G. (1978). A neurological model for childhood autism. *Archives of Neurology*, 37, 504–510.
- D'Amato, R. C. (1990). A neuropsychological approach to school psychology. *School Psychology Quarterly*, 5, 141–160.
- D'Amato, R. C., Gray, J. W., & Dean, R. S. (1988). A comparison between intelligence and neuropsychological functioning. *Journal of School Psychology*, 26, 283–292.
- D'Amato, R. C., Hammons, P. F., Terminie, T. J., & Dean, R. S. (1992). Neuropsychological training in APA-accredited and nonaccredited school psychology programs. *Journal of School Psychology*, 30, 175–183.
- D'Amato, R. C., & Rothlisberg, B. A. (1996). How education should respond to students with traumatic brain injuries. *Journal of Learning Disabilities*, 29, 670–683.
- D'Amato, R. C., Rothlisberg, B. A., & Leu, P. W. (in press). Neuropsychological assessment for intervention. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed.). New York: Wiley.
- D'Amato, R. C., Rothlisberg, B. A., & Rhodes, R. L. (1997). Utilizing neuropsychological paradigms for understanding common educational and psychological tests. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 270–295). New York: Plenum.
- Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successive cognitive processes*. New York: Academic Press.
- Dean, R. S. (1984). Functional lateralization of the brain. *Journal of Special Education*, 8, 239–256.
- Dean, R. S. (1985). Foundation and rationale for neuropsychological bases of individual differences. In

- L. D. Hartlage & C. F. Telzrow (Eds.), *The neuropsychology of individual differences: A developmental perspective* (pp. 203–244). New York: Plenum.
- Dean, R. S. (1986). Lateralization of cerebral functions. In D. Wedding, A. M. Horton, & J. S. Webster (Eds.), *The neuropsychology handbook: Behavioral and clinical perspectives* (pp. 80–102). Berlin, Germany: Springer-Verlag.
- Dean, R. S., & Gray, J. W. (1990). Traditional approaches to neuropsychological assessment. In C. R. Reynolds, & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (pp. 317–388). New York: Guilford.
- Dean, R. S., & Woodcock, R. W. (in press). *Dean Woodcock neuropsychological assessment system professional manual*. Manuscript in preparation, Ball State University.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1994). *CVLT-C Children's California Verbal Learning Test: Manual*. San Antonio, TX: Psychological Corporation.
- Denckla, M. B. (1994). Measurement of executive function. In G. R. Lyon, *Frames of reference of the assessment of learning disabilities: New views on measurement issues* (pp. 117–142). Baltimore: Brookes.
- Denckla, M. B., LeMay, M., & Chapman, C. A. (1985). Few CT scan abnormalities found even in neurologically impaired learning disabled children. *Journal of Learning Disabilities*, 18, 132–135.
- Derryberry, D., & Reed, M. A. (1996). Regulatory processes and the development of cognitive representations. *Development and Psychopathology*, 8, 215–234.
- Dietzen, S. R. (1986). *Hemispheric specialization for verbal sequential and nonverbal simultaneous information processing styles of low-income 3 to 5 year olds*. Unpublished doctoral dissertation, Washington State University.
- Donders, J. (1992). Validity of the Kaufman Assessment Battery for Children when employed with children with traumatic brain injury. *Journal of Clinical Psychology*, 48, 225–229.
- Duffy, F. H., Denckla, M. B., McAnulty, G. B., & Holmes, J. A. (1988). Neurophysiological studies in dyslexia. In F. Plum (Ed.), *Language, communication, and the brain* (pp. 105–122). New York: Raven.
- Duffy, F. H., & McAnulty, G. (1990). Neurophysiological heterogeneity and the definition of dyslexia: Preliminary evidence for plasticity. *Neuropsychologia*, 28, 555–571.
- Edmonds, J. E., Cohen, M. J., Riccio, C. A., Bacon, K. L., & Hynd, G. W. (1993, October). *The development of clock face drawing in normal children*. Paper presented at the annual meeting of the National Academy of Neuropsychology, Phoenix, AZ.
- Efron, R. (1990). *The decline and fall of hemispheric specialization*. Hillsdale, NJ: Erlbaum.
- Evans, L. P., Tannehill, R., & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. *Behavior Research Methods, Instruments, and Computers*, 27(2), 162–165.
- Fan, X., Willson, V. L., & Reynolds, C. R. (1995). Assessing the similarity of the construct structure of the KABC for black and white children from 7 to 12½ years in age. *Journal of Psychoeducational Assessment*, 13, 120–131.
- Feagans, L. V., Short, E. J., & Meltzer, L. J. (1991). *Subtypes of learning disabilities: Theoretical perspectives and research*. Hillsdale, NJ: Erlbaum.
- Fennell, E. B. (1994). Issues in child neuropsychological assessment. In R. Venderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (pp. 165–184). Hillsdale, NJ: Erlbaum.
- Fennell, E. B., & Bauer, R. M. (1997). Models of inference in evaluating brain-behavior relationships in children. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 204–215). New York: Plenum.
- First, M. B. (1994). Computer-assisted assessment of DSM III-R diagnoses. *Psychiatric Annals*, 24, 25–29.
- Fletcher, J. M. (1988). Brain-injured children. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral assessment of childhood disorders* (Vol. 2, pp. 451–589). New York: Guilford.
- Fletcher, J. M., Shaywitz, B. A., & Shaywitz, S. E. (1994). Attention as a process and as a disorder. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 103–116). Baltimore: Brookes.
- Fletcher, J. M., & Taylor, H. G. (1984). Neuropsychological approaches to children: Toward a developmental neuropsychology. *Journal of Clinical Neuropsychology*, 6, 139–156.
- Foxcroft, C. D. (1989). Factor analysis of the Reitan-Indiana Neuropsychological Test Battery. *Perceptual and Motor Skills*, 69, 1303–1313.
- Gaddes, W. H. (1980). *Learning disabilities and brain function*. Berlin, Germany: Springer-Verlag.
- Gaddes, W. H. (1983). Applied educational neuropsychology: Theories and problems. *Journal of Learning Disabilities*, 16, 511–514.
- Gaddes, W. H., & Edgell, D. (1994). *Learning disabilities and brain function: A neurodevelopmental approach*. New York: Springer-Verlag.
- Gatten, S. L., Arceneaux, J. M., Dean, R. S., & Anderson, J. L. (1994). Perinatal risk factors as predictors of developmental functioning. *International Journal of Neuroscience*, 75, 167–174.
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114, 345–362.
- Geary, D. C., & Gilger, J. W. (1984). The Luria-Nebraska Neuropsychological Battery-Children's Revision: Comparison of learning disabled and normal children matched on full scale IQ. *Perceptual and Motor Skills*, 58, 115–118.
- Golden, C. J. (1981). The Luria-Nebraska children's battery: Theory and formulation. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-age child: Issues and procedures* (pp. 277–302). New York: Grune & Stratton.
- Golden, C. J. (1984). *Luria-Nebraska neuropsychological battery: Children's revision*. Los Angeles: Western Psychological Services.
- Golden, C. J. (1997). The Nebraska neuropsychological children's battery. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 237–251). New York: Plenum.
- Goldman, P. S., & Lewis, M. E. (1978). Developmental biology of brain damage and experience. In C. W. Cotman (Ed.), *Neuronal plasticity*. New York: Raven.
- Goldman-Rakic, P. S. (1987). Development of cortical circuitry and cognitive function. *Child Development*, 58, 601.
- Goldstein, G. (1997). The clinical utility of standardized or flexible battery approaches to neuropsychological assessment. In G. Goldstein & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 67–92). New York: Plenum.
- Gonzalez, V., Brusca-Vega, R., & Yawkey, T. (1997). *Assessment and instruction of culturally and linguistically diverse students with or at-risk of learning problems*. Boston: Allyn & Bacon.
- Gordon, M. (1983). *The Gordon Diagnostic System*. DeWitt, NY: Gordon System.
- Gray, J. A. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. Oxford, UK: Oxford University Press.
- Gray, J. W., & Dean, R. S. (1990). Implications of

- neuropsychological research for school psychology. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (pp. 269–288). New York: Wiley.
- Gray, J. W., Dean, R. S., & Rattann, G. (1987). Assessment of perinatal risk factors. *Psychology in the Schools*, 24, 15–21.
- Gross-Tsur, V., Salev, R. S., Manor, O., & Amir, N. (1995). Developmental right hemisphere syndrome: Clinical spectrum of the nonverbal learning disability. *Journal of Learning Disabilities*, 28, 80–86.
- Gulbrandson, G. B. (1984). Neuropsychological sequelae of light head injuries in older children 6 months after trauma. *Journal of Clinical Neuropsychology*, 6, 257–268.
- Gutkin, T. J., & Reynolds, C. R. (1980, September). *Normative data for interpreting Reitan's index of Wechsler subtest scatter*. Paper presented at the annual meeting of the American Psychological Association, Montreal, Canada.
- Haak, R. (1989). Establishing neuropsychology in a school setting: Organization, problems, and benefits. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 489–502). New York: Plenum.
- Hale, R. L., & Foltz, S. G. (1982). Prediction of academic achievement in handicapped adolescents using a modified form of the Luria Nebraska Pathognomonic Scale and WISC-R Full Scale IQ. *Clinical Neuropsychology*, 4, 99–102.
- Hall, C. W., & Kataria, S. (1992). Effects of two treatment techniques on delay and vigilance tasks with attention deficit hyperactive disorder (ADHD) children. *Journal of Psychology*, 126, 17–25.
- Halperin, J. M. (1991). The clinical assessment of attention. *International Journal of Neuroscience*, 58, 171–182.
- Halperin, J. M., McKay, K. E., Matier, K., & Sharma, V. (1994). Attention, response inhibition, and activity level in children: Developmental neuropsychological perspectives. In M. G. Tramontana & S. R. Hooper (Eds.), *Advances in child neuropsychology* (Vol. 2., pp. 1–54). New York: Springer-Verlag.
- Harbord, M. G., Finn, J. P., Hall-Craggs, M. A., Robb, S. A., Kendall, B. E., & Boyd, S. G. (1990). Myelination patterns on magnetic resonance of children with developmental delay. *Developmental Medicine and Child Neurology*, 32, 295–303.
- Harrington, D. E. (1990). Educational strategies. In M. Rosenthal, E. R. Griffith, M. R. Bond, & J. D. Miller (Eds.), *Rehabilitation of the adult and child with traumatic brain injury* (2nd ed., pp. 476–492). Philadelphia: Davis.
- Hartlage, L. C. (1975). Neuropsychological approaches to predicting outcome of remedial education strategies for learning disabled children. *Pediatric Psychology*, 3, 23–28.
- Hartlage, L. C. (1982). Neuropsychological assessment techniques. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 296–313). New York: Wiley.
- Hartlage, L. C., & Long, C. J. (1997). Development of neuropsychology as a professional specialty: History, training, and credentialing. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 3–16). New York: Plenum.
- Hartlage, L. C. & Reynolds, C. R. (1981). Neuropsychological assessment and the individualization of instruction. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment of the school-aged child* (pp. 355–378). Boston: Allyn & Bacon.
- Hartlage, L. C., & Telzrow, C. F. (1983). The neuropsychological basis of educational intervention. *Journal of Learning Disabilities*, 16, 521–528.
- Hartlage, L. C., & Telzrow, C. F. (1986). *Neuropsychological assessment and intervention with children and adolescents*. Sarasota, FL: Professional Resource Exchange.
- Hartlage, P. L., & Givens, T. S. (1982). Common neurological problems of school age children. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 1009–1222). New York: Wiley.
- Hartlage, P. L., & Hartlage, L. C. (1978). Clinical consultation to pediatric neurology and developmental pediatrics. *Journal of Clinical Child Psychology*, 12, 52–53.
- Haut, J. S., Haut, M. W., Callahan, T. S., & Franzen, M. D. (1992, November). *Factor analysis of the Wide Range Assessment of Memory and Learning (WRAML) scores in a clinical sample*. Paper presented at the 12th Annual Meeting of the National Academy of Neuropsychology, Pittsburgh, PA.
- Heaton, R. K. (1981). *A manual for the Wisconsin Card Sorting Test*. Odessa, FL: Psychological Assessment Resources.
- Heilman, K. M., Watson, R. T., & Valenstein, E. (1985). Neglect and related disorders. In K. M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (2nd ed., pp. 243–293). New York: Oxford University Press.
- Hendren, R. L., Hodde-Vargas, J., Yeo, R. A., & Vargas, L. A. (1995). Neuropsychophysiological study of children at risk for schizophrenia: A preliminary report. *Journal of the American Academy of Child and Adolescent Psychiatry*, 34, 1284–1291.
- Hiscock, M., & Kinsbourne, M. (1987). Specialization of the cerebral hemispheres: Implications for learning. *Journal of Learning Disabilities*, 20, 130.
- Hooper, S. R., Boyd, T. A., Hynd, G. W., & Rubin, J. (1993). Definitional issues and neurobiological foundations of selected severe neurodevelopmental disorders. *Archives of Clinical Neuropsychology*, 8, 297–307.
- Hooper, S. R., & Hynd, G. W. (1985). Differential diagnosis of subtypes of developmental dyslexia with the Kaufman Assessment Battery for Children (K-ABC). *Journal of Clinical Child Psychology*, 14, 145–152.
- Hooper, S. R., & Tramontana, M. G. (1997). Advances in neuropsychological bases of child and adolescent psychopathology: Proposed models, findings, and on-going issues. *Advances in Clinical Child Psychology*, 19, 133–175.
- Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate psychology* (2nd ed., pp. 645–685). New York: Academic.
- Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *WJ-R technical manual*. Chicago: Riverside.
- Howieson, D. B., & Lezak, M. D. (1992). The neuropsychological evaluation. In S. C. Yudofsky & R. E. Hales, (Eds.), *The American Psychiatric Press textbook of neuropsychiatry* (2nd ed., pp. 127–150). Washington, DC: American Psychiatric Press.
- Hurd, A. (1996). A developmental cognitive neuropsychological approach to the assessment of information processing in autism. *Child Language Teaching and Therapy*, 12, 288–299.
- Hynd, G. W. (1981). Neuropsychology in schools. *School Psychology Review*, 10, 480–486.
- Hynd, G. W. (1992). *Neuropsychological assessment in clinical child psychology*. Newbury Park, CA: Sage.
- Hynd, G. W., & Cohen, M. J. (1983). *Dyslexia: Neuropsychological theory, research, and clinical differentiation*. New York: Grune & Stratton.
- Hynd, G. W., Marshall, R. M., & Semrud-Clikeman, M. (1991). Developmental dyslexia, neurolinguistic theory and deviations in brain morphology. *Reading and Writing: An Interdisciplinary Journal*, 3, 345–362.
- Hynd, G. W., & Willis, W. G. (1988). *Pediatric neuropsychology*. New York: Grune & Stratton.
- Iivaneihan, M., Launes, J., Pihko, H., Nikkinen, P., &

- Lindroth, L. (1990). Single photon emission computed tomography of brain perfusion: Analysis of 60 pediatric cases. *Developmental Medicine and Child Neurology*, 32, 63–68.
- Jarvis, P. E., & Barth, J. T. (1984). *Halstead-Reitan Test Battery: An interpretive guide*. Odessa, FL: PAR.
- Jernigan, T. L., & Tallal, P. (1990). Late childhood changes in brain morphology observable with MRI. *Developmental Medicine and Child Neurology*, 32, 379–385.
- Kail, R. (1984). *The development of memory in children*. San Francisco: Freeman.
- Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence*. Boston: Allyn & Bacon.
- Kamphaus, R. W., & Reynolds, C. R. (1987). *Clinical and research applications of the K-ABC*. Circle Pines, MN: American Guidance Service.
- Kane, R. L., & Kay, G. G. (1997). Computer applications in neuropsychological assessment. In G. Goldstein & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 359–392). New York: Plenum.
- Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function* (pp. 125–167). Washington, DC: American Psychological Association.
- Karras, D., Newton, D. B., Franzen, M. D., & Golden, C. J. (1987). Development of factor scales for Luria-Nebraska Neuropsychological Battery: Children's revision. *Journal of Clinical Child Psychology*, 16, 19–28.
- Kaufman, A. S. (1976a). A new approach to the interpretation of test scatter on the WISC-R. *Journal of Learning Disabilities*, 9, 160–167.
- Kaufman, A. S. (1976b). Verbal-performance IQ discrepancies on the WISC-R. *Journal of Learning Disabilities*, 9, 739–744.
- Kaufman, A. S. (1979). Cerebral specialization and intelligence testing. *Journal of Research and Development in Education*, 12, 96–107.
- Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children (K-ABC) administration and scoring manual*. Circle Pines, MN: American Guidance Services.
- Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children (K-ABC) interpretative manual*. Circle Pines, MN: American Guidance Services.
- Kinsbourne, M. (1975). Cerebral dominance, learning, and cognition. In H. R. Myklebust (Ed.), *Progress in learning disabilities*. New York: Grune & Stratton.
- Kinsbourne, M. (1989). A model of adaptive behavior related to cerebral participation in emotional control. In G. Gainotti & C. Caltagirone (Eds.), *Emotions and the dual brain* (pp. 248–260). New York: Springer-Verlag.
- Klesges, R. C. (1983). The relationship between neuropsychological, cognitive, and behavioral assessments of brain functioning in children. *Clinical Neuropsychology*, 5, 28–32.
- Klonoff, H., & Low, M. (1974). Disordered brain function in young children and early adolescents: Neuropsychological and electroencephalographic correlates. In R. M. Reitan & L. A. Davison (Eds.), *Clinical neuropsychology: Current status and application* (pp. 76–94). Washington, DC: Winston.
- Knights, R. M., & Norwood, J. W. (1979). *A neuropsychological test battery for children: Examiner's manual*. Ottawa, Canada: Knights Psychological Consultants.
- Koriath, U., Gualtieri, C. T., van Bourgondien, M. E., Quade, D., & Werry, J. S. (1985). Construct validity of clinical diagnosis in pediatric psychiatry: Relationship among measures. *Journal of the American Academy of Child Psychiatry*, 24, 429–436.
- Korkman, M. (1988). NEPSY: An adaptation of Luria's investigation for young children. *Clinical Neuropsychologist*, 2, 375–392.
- Korkman, M., & Hakkinen-Rihu, P. (1994). A new classification of developmental language disorders. *Brain & Language*, 47(1), 96–116.
- Korkman, M., Kirk, U., & Kemp, S. (1997). *The neuropsychological investigation for children*. San Antonio, TX: Psychological Corporation.
- Korkman, M., Liikanen, A., & Fellman, V. (1996). Neuropsychological consequences of very low birth weight and asphyxia at term: Follow-up until school age. *Journal of Clinical Neuropsychology*, 18, 220–233.
- Leark, R. A., Snyder, T., Grove, T., & Golden, C. J. (1983, August). *Comparison of the KABC and standardized neuropsychological batteries: Preliminary results*. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.
- Leu, P. W., & D'Amato, R. C. (1994, April). *Right children, wrong teachers? Using an ecological assessment for placement decisions*. Paper presented at the 26th Annual Convention of the National Association of School Psychologists, Seattle, WA.
- Levin, H. S., Culhane, K. A., Hartmann, J., Evankovich, K., Mattson, A. J., Harward, H., Ringholz, G., Ewing-Cobbs, L., & Fletcher, J. M. (1991). Developmental changes in performance on tests of purported frontal lobe functioning. *Developmental Neuropsychology*, 7, 377–395.
- Levine, M. D. (1993). *Developmental variation and learning disorders*. Cambridge, MA: Education Publishers Service.
- Lezak, M. D. (1995). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.
- Little, S. G., & Stavrou, E. (1993). The utility of neuropsychological approaches with children. *The Behavior Therapist*, 16, 104–106.
- Livingston, R. B., Pritchard, D. A., Moses, J. A., Haak, R. A., Marshall, R., & Gray, R. (1997). Modal profiles for the Halstead-Reitan neuropsychological battery for children. *Archives of Clinical Neuropsychology*, 12, 450–476.
- Lovrich, D., Cheng, J. C., & Velting, D. M. (1996). Late cognitive brain potentials, phonological and semantic classification of spoken words, and reading ability in children. *Journal of Clinical Neuropsychology*, 18, 161–177.
- Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.
- Luria, A. R. (1970). Functional organization of the brain. *Scientific American*, 222, 66–78.
- Luria, A. R., (1973). *The working brain*. New York: Basic Books.
- Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.
- Majovski, L. V. (1984). The K-ABC: Theory and applications for child neuropsychological assessment and research. *Journal of Special Education*, 18, 266–268.
- Mataszow, G., & Hynd, G. W. (1992, February). *Analysis of the anterior-posterior gradient hypothesis as applied to attention deficit disorder children*. Paper presented at the annual meeting of the International Neuropsychological Society, San Diego, CA.
- Mattarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins.
- Maurer, R. G., & Damasio, A. R. (1982). Childhood autism from the point of view of behavioral neurology. *Journal of Autism and Developmental Disorders*, 12, 195–205.
- Mayfield, J. W., & Reynolds, C. R. (1997). Black-white differences in memory test performance among children and adolescents. *Archives of Clinical Neuropsychology*, 12, 111–122.
- McBurnett, K., Hynd, G. W., Lahey, B. B., & Town, P. A. (1988). Do neuropsychological measures contribute to the prediction of academic achievement? The predictive

- validity of the LNNB-CR pathognomonic scale. *Journal of Psychoeducational Assessment*, 6, 162–167.
- McGlone, J., & Davidson, W. (1973). The relation between spatial ability with special reference to sex and hand preference. *Neuropsychologia*, 11, 105–113.
- Merola, J. L., & Leiderman, J. (1985). The effect of task difficulty upon the extent to which performance benefits from between hemisphere division of inputs. *International Journal of Neuroscience*, 51, 35–44.
- Mesulam, M. M. (1985). *Principles of behavioral neurology*. Philadelphia: F. A. Davis.
- Milberg, W. B., Hebben, N., & Kaplan, E. (1986). The Boston process approach to neuropsychological assessment. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment and neuropsychiatric disorders* (2nd ed., pp. 58–80). New York: Oxford University Press.
- Miller, L. T., & Vernon, P. A. (1996). Intelligence, reaction time, and working memory in 4- to 6-year-old children. *Intelligence*, 22, 155–190.
- Mitchell, W. G., Chavez, J. M., Baker, S. A., Guzman, B. L., & Azen, S. P. (1990). Reaction time, impulsivity, and attention in hyperactive children and controls: A video game technique. *Journal of Child Neurology*, 5, 195–204.
- Moffitt, T. E. (1993). The neuropsychology of conduct disorder. *Developmental Psychopathology*, 5, 135–151.
- Molfese, D. L. (1995). Electrophysiological responses obtained during infancy and their relation to later language development: Further findings. In M. G. Tramontana & S. R. Hooper, (Eds.), *Advances in Child Neuropsychology* (Vol. 3, pp. 1–11). New York: Springer-Verlag.
- Morgan, S. B., & Brown, T. L. (1988). Luria-Nebraska Neuropsychological Battery-Children's Revision: Concurrent validity with three learning disability subtypes. *Journal of Consulting and Clinical Psychology*, 56, 463–466.
- Morris, J. M., & Bigler, E. (1985, January). *An investigation of the Kaufman Assessment Battery for Children (KABC) with neurologically impaired children*. Paper presented at the annual meeting of the International Neuropsychological Society, San Diego, CA.
- Morris, R. (1994). Multidimensional neuropsychological assessment models. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement* (pp. 515–522). Baltimore: Brookes.
- Novak, G. P., Solanto, M., & Abikoff, H. (1995). Spatial orienting and focused attention in attention deficit hyperactivity disorder. *Psychophysiology*, 32, 546–559.
- Nussbaum, N. L., & Bigler, E. D. (1986). Neuropsychological and behavioral profiles of empirically derived subgroups of learning disabled children. *International Journal of Clinical Neuropsychology*, 8, 82–89.
- Nussbaum, N. L., & Bigler, E. D. (1990). *Identification and treatment of attention deficit disorder*. Austin, TX: Pro-Ed.
- Nussbaum, N. L., & Bigler, E. D. (1997). Halstead-Reitan neuropsychological test batteries for children. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 219–236). New York: Plenum.
- Nussbaum, N. L., Bigler, E. D., Koch, W. R., Ingram, J. W., Rosa, L., & Massman, P. (1988). Personality/behavioral characteristics in children: Differential effects of putative anterior versus posterior cerebral asymmetry. *Archives of Clinical Neuropsychology*, 3, 127–135.
- Obrzut, J. E. (1981). Neuropsychological procedures with school-age children. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-age child: Issues and procedures* (pp. 237–275). New York: Grune & Stratton.
- Obrzut, J. E., & Hynd, G. W. (1983). The neurobiological and neuropsychological foundations of learning disabilities. *Journal of Learning Disabilities*, 16, 515–520.
- Obrzut, J. E., & Hynd, G. W. (1986). Child neuropsychology: An introduction to theory and research. In G. W. Hynd & J. E. Obrzut (Eds.), *Child neuropsychology* (Vol. 1, pp. 1–12). New York: Academic Press.
- Oehler-Stinnett, J., Stinnett, T. A., Wesley, A. L., & Anderson, H. N. (1988). The Luria Nebraska Neuropsychological Battery-Children's Revision: Discrimination between learning disabled and slow learner children. *Journal of Psychoeducational Assessment*, 6, 24–34.
- Parsons, O. A., & Prigatano, G. P. (1978). Methodological considerations in clinical neuropsychological research. *Journal of Consulting and Clinical Psychology*, 46, 608–619.
- Passler, M., Isaac, W., & Hynd, G. W. (1985). Neuropsychological development of behavior attributed to frontal lobe functioning in children. *Developmental Neuropsychology*, 1, 349–370.
- Pedhazur, E. J. (1973). *Multiple regression in behavioral research: Explanation and prediction*. New York: CBS College Publishing.
- Pfeiffer, S. I., Naglieri, J. A., & Tingstrom, D. H. (1987). Comparison of the Luria Nebraska Neuropsychological Battery-Children's Revision and the WISC-R with learning disabled children. *Perceptual and Motor Skills*, 65, 911–916.
- Phelps, L. (1995). Exploratory factor analysis of the WRAML with academically at-risk students. *Journal of Psychoeducational Assessment*, 13, 384–390.
- Phelps, L. (1996). Discriminative validity of the WRAML with ADHD and LD children. *Psychology in the Schools*, 33, 5–12.
- Plaisted, J. R., Gustavson, J. C., Wilkening G. N., & Golden, C. J. (1983). The Luria Nebraska Neuropsychological Battery-Children's Revision: Theory and current research findings. *Journal of Clinical Child Psychology*, 12, 13–21.
- Powell, H. (1997). Comment on computerized assessment of arithmetic computation skills with MicroCog. *Journal of the International Neuropsychological Society*, 3, 200.
- Powell, D. H., Kamplian, E. F., Thitla, D., Weintraub, S., Catlin, R., & Funkenstein, H. H. (1993). *MicroCog assessment of cognitive functioning manual*. San Antonio, TX: Psychological Corporation.
- Ramsey, M. C., & Reynolds, C. R. (1995). Separate digits tests: A brief history, a literature review, and reexamination of the factor structure of the Test of Memory and Learning (TOMAL). *Neuropsychology Review*, 5, 151–171.
- Reitan, R. M. (1969). *Manual for the administration of neuropsychological test batteries for adults and children*. Indianapolis, IN: Author.
- Reitan, R. M. (1974). *Clinical neuropsychology: Current status and applications*. New York: Winston.
- Reitan, R. M. (1986). *Theoretical and methodological bases of the Halstead-Reitan Neuropsychological Test Battery*. Tucson, AZ: Neuropsychological Press.
- Reitan, R. M. (1987). *Neuropsychological evaluation of children*. Tucson, AZ: Neuropsychological Press.
- Reitan, R. M., & Davison, L. A. (1974). *Clinical neuropsychology: Current status and applications*. Washington, DC: Winston.
- Reitan, R. M., & Wolfson, D. (1985). *The Halstead Reitan neuropsychological battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychological Press.
- Reitan, R. M., & Wolfson, D. (1988). The Halstead Reitan Neuropsychological Test Battery and REHABIT: A model for integrating evaluation and remediation of cognitive impairment. *Cognitive Rehabilitation*, May-June, 10–17.
- Reschly, D., & Gresham, F. M. (1989). Current neuropsychological diagnosis of learning problems: A leap of faith. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 503–520). New York: Plenum.
- Reynolds, C. R. (1979). Interpreting the index of abnormal

- ality when the distribution of score differences is known: Comment on Piotrowski. *Journal of Consulting and Clinical Psychology*, 47, 401–402.
- Reynolds, C. R. (1981a). The neuropsychological basis of intelligence. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-aged child: Issues and procedures* (pp. 87–124). New York: Grune & Stratton.
- Reynolds, C. R. (1981b). Neuropsychological assessment and the habilitation of learning: Considerations in the search for the aptitude $\times$ treatment interaction. *School Psychology Review*, 10, 343–349.
- Reynolds, C. R. (1982). The importance of norms and other traditional psychometric concepts to assessment in clinical neuropsychology. In R. N. Malathesha & L. C. Hartlage (Eds.), *Neuropsychology and cognition* (Vol. 3, pp. 55–76). The Hague, The Netherlands: Nijhoff.
- Reynolds, C. R. (1986a). Transactional models of intellectual development, yes. Deficit models of process remediation, no. *School Psychology Review*, 15, 256–260.
- Reynolds, C. R. (1986b). Clinical acumen but psychometric naivete in neuropsychological assessment of educational disorders. *Archives of Clinical Neuropsychology*, 1(2), 121–137.
- Reynolds, C. R. (1992). Two key concepts in the diagnosis of learning disabilities and the habilitation of learning. *Learning Disability Quarterly*, 15(1), 2–12.
- Reynolds, C. R. (1997a). Forward and backward memory span should not be combined for clinical analysis. *Archives of Clinical Neuropsychology*, 12, 29–40.
- Reynolds, C. R. (1997b). Measurement and statistical problems in neuropsychological assessment of children. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 180–203). New York: Plenum.
- Reynolds, C. R. (1997c). Postscripts on premorbid ability estimation: Conceptual addenda and a few words on alternative and conditional approaches. *Archives of Clinical Neuropsychology*, 12, 769–778.
- Reynolds, C. R., & Bigler, E. D. (1994). *Manual for the Test of Memory and Learning*. Austin, TX: PRO-ED.
- Reynolds, C. R., & Bigler, E. D. (1996). Factor structure, factor indexes, and other useful statistics for interpretation of the Test of Memory and Learning (TOMAL). *Archives of Clinical Neuropsychology*, 11, 29–43.
- Reynolds, C. R., & Bigler, E. D. (1997a). Clinical neuropsychological assessment of child and adolescent memory with the Test of Memory and Learning. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 296–319). New York: Plenum.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Services.
- Reynolds, C. R., & Kamphaus, R. W. (1997). The Kaufman assessment battery for children: Development, structure and applications in neuropsychology. In A. M. Horton, D. Wedding, & J. Webster, (Eds.), *The neuropsychology handbook* (Vol. 1, pp. 291–330). New York: Springer.
- Reynolds, C. R., Kamphaus, R. W., Rosenthal, B. L., & Hiemenz, J. R. (1997). Application of the Kaufman assessment battery for children (K-ABC) in neuropsychological assessment. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 252–269). New York: Plenum.
- Reynolds, C. R., Wilen, S., & Stone, B. (1997, November). *The economy of neuropsychological evaluations*. Paper presented at the annual meeting of the National Academy of Neuropsychology, Las Vegas, NV.
- Riccio, C. A., Gonzalez, J. J., & Hynd, G. W. (1994). Attention-deficit hyperactivity disorder (ADHD) and learning disabilities. *Learning Disability Quarterly*, 17, 311–322.
- Riccio, C. A., Hall, J., Morgan, A., Hynd, G. W., Gonzalez, J. J., & Marshall, R. M. (1994). Executive function and the Wisconsin card sorting test: Relationship with behavioral ratings and cognitive ability. *Developmental Neuropsychology*, 10, 215–229.
- Riccio, C. A., & Hynd, G. W. (1995). Contributions of neuropsychology to our understanding of developmental reading problems. *School Psychology Review*, 24, 415–425.
- Riccio, C. A., & Hynd, G. W. (1996). Neuroanatomical and neurophysiological aspects of dyslexia. *Topics in Language Disorders*, 16(2), 1–13.
- Riccio, C. A., & Hynd, G. W., & Cohen, M. J. (1993). Neuropsychology in the schools: Does it belong? *School Psychology International*, 14, 291–315.
- Riccio, C. A., Hynd, G. W., & Cohen, M. J. (1996). Etiology and neurobiology of Attention-Deficit Hyperactivity Disorder. In W. Bender (Ed.), *Understanding ADHD: A practical guide for teachers and parents* (pp. 23–44). New York: Merrill.
- Ris, M. D., & Noll, R. B. (1994). Long-term neurobehavioral outcome in pediatric brain tumor patients: Review and methodological critique. *Journal of Clinical and Experimental Neuropsychology*, 16, 21.
- Rothlisberg, B. A., & D'Amato, R. C. (1988). Increased neuropsychological understanding seen as important for school psychologists. *Communique*, 17(2), 4–5.
- Rourke, B. P. (1984). *Subtype analysis of learning disabilities*. New York: Guilford.
- Rourke, B. P. (1989). *Nonverbal learning disabilities: The syndrome and the model*. New York: Guilford.
- Rourke, B. P. (1991). *Neuropsychological validation of learning disability subtypes*. New York: Guilford.
- Rourke, B. P. (1994). Neuropsychological assessment of children with learning disabilities: Measurement issues. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 475–514). Baltimore: Brookes.
- Rourke, B. P., Bakker, D. J., Fisk, J. L., & Strang, J. D. (1983). *Child neuropsychology: An introduction to theory, research, and practice*. New York: Guilford.
- Rourke, B. P., Fisk, J. L., & Strang, J. D. (1986). *Neuropsychological assessment of children: A treatment oriented approach*. New York: Guilford.
- Rutter, M. (1981). Psychological sequelae of brain damage in children. *American Journal of Clinical Neuropsychology*, 138, 1533–1544.
- Rutter, M. (1983). *Developmental neuropsychiatry*. New York: Guilford.
- Rutter, M., Graham, P., & Yule, W. (1970). *A neuropsychiatric study in childhood*. London: Lavenham Press.
- Saigal, S. (1995). Long term outcome of very low-birth-weight infants: Kindergarten and beyond. *Developmental Brain Dysfunction*, 8, 109–118.
- Samuels, S. J. (1979). An outside view of neuropsychological testing. *Journal of Special Education*, 13, 57–60.
- Sandoval, J. (1981, August). *Can neuropsychology contribute to rehabilitation in educational settings? No*. Paper presented at the annual meeting of the American Psychological Association, Los Angeles, CA.
- Sandoval, J., & Halperin, R. M. (1981). A critical commentary on neuropsychology in the schools: Are we ready? *School Psychology Review*, 10, 381–388.
- Segalowitz, S. (1983). *Language functions and brain organization*. New York: Academic Press.
- Seidel, U. P., Chadwick, O., & Rutter, M. (1975). Psychological disorders in crippled children: A comparative study of children with and without brain damage. *Developmental Medicine and Child Neurology*, 17, 563.
- Seidman, L. J., Biederman, J., Faraone, S. V., Milberger, S., Norman, D., Seiverd, K., Benedict, K., Guite, J.,

- Mick, E., & Kiely, K. (1995). Effects of family history and comorbidity on the neuropsychological performance of children with ADHD: Preliminary findings. *Journal of the American Academy of Child and Adolescent Psychiatry*, 34, 1015–1024.
- Selz, M. (1981). Halstead-Reitan neuropsychological test batteries for children. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment of the school-aged child: Issues and procedures*, (pp. 195–235). New York: Grune & Stratton.
- Selz, M., & Reitan, R. M. (1979a). Rules for neuropsychological diagnosis: Classification of brain function in older children. *Journal of Consulting and Clinical Psychology*, 47, 258–264.
- Selz, M., & Reitan, R. M. (1979b). Neuropsychological test performance of normal, learning disabled, and brain damaged older children. *Journal of Nervous and Mental Disease*, 167, 298–302.
- Shapiro, E. G., & Dotan, N. (1985, October). *Neurological findings and the Kaufman Assessment Battery for Children*. Paper presented at the annual meeting of the National Association of Neuropsychologists, Philadelphia, PA.
- Sheslow, D., & Adams, W. (1990). *Wide Range Assessment of Memory and Learning*. Wilmington, DE: Jastak Associates.
- Shields, J., Varley, R., Brooks, P., & Simpson, A. (1996). Hemisphere function in developmental language disorders and high level autism. *Developmental Medicine & Child Neurology*, 38, 473–486.
- Shurtleff, H. A., Fay, G. E., Abbott, R. D., & Berninger, V. W. (1988). Cognitive and neuropsychological correlates of academic achievement: A levels of analysis assessment model. *Journal of Psychoeducational Assessment*, 6, 298–308.
- Snow, J. H., & Hooper, S. R. (1994). *Pediatric traumatic brain injury*. Thousand Oaks, CA: Sage.
- Snow, J. H., & Hynd, G. W. (1985a). Factor structure of the Luria-Nebraska Neuropsychological Battery-Children's Revision. *Journal of School Psychology*, 23, 271–276.
- Snow, J. H., & Hynd, G. W. (1985b). A multivariate investigation of the Luria-Nebraska Neuropsychological Battery-Children's Revision with learning disabled children. *Journal of Psychoeducational Assessment*, 2, 23–28.
- Snow, J. H., Hynd, G. W., & Hartlage, L. H. (1984). Differences between mildly and more severely learning disabled children on the Luria Nebraska Neuropsychological Battery-Children's Revision. *Journal of Psychoeducational Assessment*, 2, 23–28.
- Snyder, T. J., Lemark, R. A., Golden, C. J., Grove, T., & Allison, R. (1983, March). *Correlations of the K-ABC, WISC-R, and Luria Nebraska Children's Battery for exceptional children*. Paper presented at the annual meeting of the National Association of School Psychologists, Detroit, MI.
- Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist*, 42, 137–144.
- Sperry, R. W. (1968). Hemisphere deconnection and unity in conscious awareness. *American Psychologist*, 23, 723–733.
- Sperry, R. W. (1974). Lateral specialization in the surgically separated hemispheres. In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program*. Cambridge, MA: MIT Press.
- Sprien, O., & Gaddes, W. H. (1979). Developmental norms for 15 neuropsychological tests age 6–15. *Cortex*, 5, 813–818.
- Sprien, O., Risser, A. H., & Edgell, D. (1995). *Developmental neuropsychology*. London: Oxford University Press.
- Strom, D. A., Gray, J. W., Dean, R. S., & Fischer, W. E. (1987). Incremental validity of the Halstead-Reitan neuropsychological battery in predicting achievement for learning disabled children. *Journal of Psychoeducational Assessment*, 5, 157–165.
- Sweet, J. J., Carr, M. A., Rossini, E., & Kasper, C. (1986). Relationship between the Luria Nebraska Neuropsychological Battery-Children's Revision and the WISC-R: Further examination using Kaufman's factors. *International Journal of Clinical Neuropsychology*, 8, 177–180.
- Sweet, J. J., & Moberg, P. (1990). A survey of practices and beliefs among ABPP and non-ABPP clinical neuropsychologists. *The Clinical Neuropsychologist*, 4, 101–120.
- Sweet, J. J., Moberg, P., & Westergaard, C. K. (1996). Five year follow-up survey of practices and beliefs of clinical neuropsychologists. *The Clinical Neuropsychologist*, 10, 202–221.
- Talley, J. L. (1986). Memory in learning disabled children: Digit span and the Rey Auditory Verbal Learning Test. *Archives of Clinical Neuropsychology*, 1, 315–322.
- Taylor, H. G. (1988). Neuropsychological testing: Relevance for assessing children's learning disabilities. *Journal of Consulting and Clinical Psychology*, 56, 795–800.
- Taylor, H. G., Barry, C. T., & Schatschneider, C. W. (1993). School-age consequences of haemophilus influenzae type b meningitis. *Journal of Clinical Child Psychology*, 22, 196–206.
- Taylor, H. G., & Fletcher, J. M. (1990). Neuropsychological assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of neuropsychological assessment* (pp. 228–255). New York: Plenum.
- Teeter, P. A. (1986). Standard neuropsychological batteries for children. In G. W. Hynd & J. E. Obrzut (Eds.), *Child neuropsychology* (Vol. 2, pp. 187–227). New York: Academic Press.
- Teeter, P. A. (1997). Neurocognitive interventions for childhood and adolescent disorders: A transactional model. In C. R. Reynolds & E. Fletcher-Janzen, *Handbook of clinical child neuropsychology* (2nd ed., pp. 387–417). New York: Plenum.
- Teeter, P. A., & Semrud-Clikemen, M. (1997). *Child neuropsychology: Assessment and interventions for neurodevelopmental disorders*. Boston: Allyn & Bacon.
- Telzow, C. F., Century, E., Harris, B., & Redmond, C. (1985, April). *Relationship between neuropsychological processing models and dyslexia subtypes*. Paper presented at the annual meeting of the National Association of School Psychologists, Las Vegas, NV.
- Temple, C. M. (1997). Cognitive neuropsychology and its application to children. *Journal of Child Psychology, Psychiatry, and Allied Disciplines*, 38, 27–52.
- Timmermans, S. R., & Christensen, B. (1991). The measurement of attention deficit in TBI children and adolescents. *Cognitive Rehabilitation*, 9, 26.
- Torgesen, J. K. (1994). Issues in the assessment of executive function: An information processing perspective. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 143–162). Baltimore: Brookes.
- Torkelson, R. D., Leibbrook, L. G., Gustavson, J. L., & Sundell, R. R. (1985). Neurological and neuropsychological effects of cerebral spinal fluid shunting in children with assumed arrested ("normal pressure") hydrocephalus. *Journal of Neurology, Neurosurgery, and Psychiatry*, 48, 799–806.
- Tramontana, M. G. (1983). Neuropsychological evaluation in child/adolescent psychiatric disorders: Current status. *Psychiatric Hospital*, 14, 158–162.
- Tramontana, M., & Hooper, S. (Eds.) (1987). *Neuropsychological assessment with children*. New York: Plenum.
- Tramontana, M., & Hooper, S. (1997). Neuropsychology of child psychopathology. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child*

- neuropsychology (2nd ed., pp. 120–139). New York: Plenum.
- Tramontana, M. G., Hooper, S. R., Curley, A. S., & Nardolillo, E. M. (1990). Determinants of academic achievement in children with psychiatric disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, 29, 265–268.
- Tucker, D. M. (1989). Neural substrates of thought and affective disorders. In G. Gainotti & C. Caltagirone (Eds.), *Emotions and the dual brain* (pp. 225–234). New York: Springer-Verlag.
- Turkheimer, E., Yeo, R. A., Jones, C., & Bigler, E. D. (1990). Quantitative assessment of covariation between neuropsychological function and location of naturally occurring lesions in humans. *Journal of Clinical and Experimental Neuropsychology*, 12, 549–565.
- Voeller, K. K. S. (1995). Clinical neurologic aspects of the right hemisphere deficit syndrome. *Journal of Child Neurology*, 10, 516–522.
- Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.
- Waber, D. P., & McCormick, M. C. (1995). Late neuropsychological outcomes in preterm infants of normal IQ: Selective vulnerability of the visual system. *Journal of Pediatric Psychology*, 20, 721–735.
- Wasserman, J. (1995, February). *Assessment and remediation of memory deficits in children*. Paper presented at the meeting of the Supervisors of School Psychologists for the New York Board of Education. New York.
- Wechsler, D. (1974). *Wechsler Intelligence Scale for Children-Revised*. New York: Psychological Corporation.
- Welsh, M. C., Pennington, B. F., & Grossier, D. B. (1991). A normative developmental study of executive function: A window on prefrontal function in children. *Developmental Neuropsychology*, 7, 131–139.
- Wherry, J. N., Paal, N., Jolly, J. B., Adam, B., Holloway, C., Everett, B., & Vaught, L. (1993). Concurrent and discriminant validity of the Gordon Diagnostic System: A preliminary study. *Psychology in the Schools*, 30, 29–36.
- Whitten, C. J., D'Amato, R. C., & Chitooran, M. M. (1992). The neuropsychological approach to interventions. In R. C. D'Amato & B. A. Rothlisberg (Eds.), *Psychological perspectives on intervention: A case study approach to prescriptions for change* (pp. 112–136). New York: Longman.
- Williams, M. A., & Boll, T. J. (1997). Recent advances in neuropsychological assessment of children. In G. Goldstein & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 231–267). New York: Plenum.
- Willson, V. L., & Reynolds, C. R. (1982). Methodological and statistical problems in determining membership in clinical populations. *Clinical Neuropsychology*, 4, 134–138.
- Wittelson, S. F. (1977). Early hemisphere specialization and interhemispheric plasticity: An empirical and theoretical review. In S. Segalowitz & F. A. Gruber (Eds.), *Language development and neurological theory* (pp. 213–287). New York: Academic Press.
- Woody, R. H. (1997). Psycholegal issues for clinical child neuropsychology. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 712–725). New York: Plenum.
- Ylvisaker, M., Choraży, A. J. L., Cohen, S. B., Mastrilli, J. P., Molitor, C. B., Nelson, J., Szekeres, S. F., Valko, A. S., & Jaffe, K. M. (1990). Rehabilitative assessment following head injury in children. In M. Rosenthal, E. R. Griffith, M. R. Bond, & J. D. Miller (Eds.), *Rehabilitation of the adult and child with traumatic brain injury* (2nd ed., pp. 521–538). Philadelphia: Davis.
- Zurcher, R. (1995). Memory and learning assessment: Missing from the learning disabilities identification process for too long. *LD Forum*, 21, 27–30.

4.11

Neuropsychological Assessment of Adults

C. MUNRO CULLUM

*University of Texas Southwestern Medical Center at Dallas,
TX, USA*

4.11.1 INTRODUCTION	304
4.11.2 APPROACHES TO ASSESSMENT IN CLINICAL NEUROPSYCHOLOGY	305
4.11.2.1 <i>The Standard Battery</i>	305
4.11.2.2 <i>Test Batteries for Specific Populations</i>	308
4.11.2.3 <i>The Hypothesis-driven Approach</i>	308
4.11.2.4 <i>Quantitative and Qualitative Examinations of Neurobehavioral Competence</i>	309
4.11.2.5 <i>Other Issues in Test Interpretation</i>	309
4.11.2.5.1 <i>Test cutoff scores</i>	310
4.11.2.5.2 <i>T-scores</i>	310
4.11.2.6 <i>Computer Interpretation of Neuropsychological Assessment Results</i>	310
4.11.2.7 <i>Cognitive Screening</i>	311
4.11.3 METHODS OF NEUROPSYCHOLOGY	311
4.11.3.1 <i>Clinical Interview and Background Information</i>	312
4.11.3.2 <i>Neuropsychological Measurement of Brain Function</i>	312
4.11.4 PRINCIPAL COGNITIVE DOMAINS FOR ASSESSMENT	315
4.11.4.1 <i>Global Cognitive/Intellectual Functioning</i>	315
4.11.4.2 <i>Academic Achievement</i>	322
4.11.4.3 <i>Executive Functioning, Problem-solving, and Reasoning</i>	323
4.11.4.3.1 <i>Tests of reasoning and problem-solving</i>	323
4.11.4.4 <i>Arousal and Orientation</i>	324
4.11.4.5 <i>Attention/Concentration</i>	325
4.11.4.5.1 <i>Assessment of attention</i>	325
4.11.4.6 <i>Language</i>	326
4.11.4.6.1 <i>Global language assessment</i>	327
4.11.4.7 <i>Visuospatial Skills</i>	328
4.11.4.7.1 <i>Visuospatial tasks</i>	329
4.11.4.8 <i>Memory</i>	332
4.11.4.8.1 <i>Clinical assessment of learning and memory</i>	332
4.11.4.8.2 <i>Clinical memory tests</i>	335
4.11.4.8.3 <i>Memory batteries</i>	335
4.11.4.8.4 <i>Verbal memory tests</i>	336
4.11.4.8.5 <i>Nonverbal memory tests</i>	337
4.11.4.9 <i>Motor and Sensory Function</i>	338
4.11.4.9.1 <i>Psychometric measures of motor and sensory function</i>	338
4.11.4.10 <i>Assessment of Motivation</i>	339
4.11.4.11 <i>Personality and Emotional Functioning</i>	340
4.11.4.12 <i>Test Selection Issues</i>	341
4.11.4.13 <i>Relationships Between Neuropsychometry and the Behavioral Geography of the Brain</i>	341

4.11.4.14 Training in Neuropsychology	342
4.11.4.15 Challenges for Neuropsychological Assessment	343
4.11.5 REFERENCES	343

4.11.1 INTRODUCTION

Clinical neuropsychology represents one of the most rapidly growing areas within the field of psychology and was one of the first to gain specialty status by the American Psychological Association in 1996. At a basic level, the neuropsychological examination represents a combination of the traditional behavioral neurologic examination with a psychometric approach to the evaluation of brain-behavior relationships. From the broader field of psychology, neuropsychology derived its emphasis in the application of psychometric procedures to quantify behavior. From its other parent discipline of neurology came the interest in evaluating brain function. The term "neuropsychology" originally evolved in the 1930s and 1940s and its popularization is often attributed to Hans-Lukas Teuber. Many early neuropsychological procedures were developed during war time to assess cognitive status and suitability of individuals for special military service. Subsequently, penetrating missile wounds to the brain became the focus of localization studies. Accordingly, many tests were created with the goal of being sensitive to focal brain insults. Other measures were developed to assess for "organicity," a now archaic term that had been used to grossly refer to brain damage or neurological deficit. Prior to the advent of modern neuroimaging procedures, neuropsychological techniques emerged as a front-line assessment procedure for the identification and localization of acquired cerebral damage. While neuropsychological techniques continue to provide this aspect of neurodiagnostic assessment to some extent, more commonly the procedures are used to describe and quantify behavior. The results of these evaluations can be used to infer cerebral integrity vs. dysfunction, to delineate cognitive strengths and weaknesses, and to assist in differential diagnosis. Assisting other professionals in differential diagnostic situations and documenting level of impairment can be a primary role for these evaluations. Other goals of neuropsychological assessment include making treatment and rehabilitation recommendations, assisting in placement issues, and in evaluating treatment response (e.g., in cases of neurosurgical or pharmacological intervention).

The neuropsychological examination can also provide useful information to patients

regarding their functioning, and feedback may be used therapeutically in terms of adjustment issues. In many cases, informing the patient that their cognitive difficulties can be documented and explained (or at least are not unexpected) can be most reassuring. Discussing the nature of a disorder and the various cognitive and emotional sequelae it may have, as well as the typical course of recovery or changes that can be expected, is also helpful in terms of setting realistic expectations and goals. Furthermore, presenting information and potential compensatory strategies or interventions that are helpful given a patient's particular situation can be rewarding. In some medical settings in particular, the neuropsychologist may play a central role in helping patients understand the nature of their difficulties by providing this information in understandable terms within an emotionally supportive context.

Among existing neurodiagnostic procedures, the neuropsychological evaluation remains the most sensitive means of assessing human brain function. Knowing the presence and location of a given lesion, for example, provides only limited information about an individual's functioning. For example, the patient in Figure 1 sustained an infarction of the right posterior cerebral artery.

How such a patient might be functioning in his or her daily life, however, remains a question that simple structural neuroimaging cannot address. For example, the individual depicted in Figure 1 might be expected to show a contralateral visual field cut (hemianopsia) based on the location of the lesion, although the impact of this as well as any other associated cognitive processing difficulties would be unknown without clinical examination.

Another example is provided in the case of the 65-year-old patient whose magnetic resonance imaging (MRI) scan is presented in Figure 2. This individual demonstrated grossly normal brain structure, with no major evidence of atrophy or particular neuropathology. Despite the normal appearance of her brain, a severe level of dementia was observed as reflected by her score of 10/30 on the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975). Although dementia is often associated with cortical atrophy, clinicians should keep in mind that dementia is a clinical diagnosis and the degree of atrophy on neuroimaging shows only a modest association



Figure 1 CT scan depicting posterior right hemisphere lesion (note that right is depicted on the left).

with level of cognitive impairment (Naugle, Cullum, Bigler, & Massman, 1986).

Given the utility of neuropsychological techniques in terms of documenting and understanding the neurobehavioral sequelae of cerebral dysfunction, a discussion of some of the major approaches and common measures used in clinical neuropsychology is in order.

4.11.2 APPROACHES TO ASSESSMENT IN CLINICAL NEUROPSYCHOLOGY

Contemporary clinical neuropsychology owes a debt to many of its early founders, and because of the youth of the field, many of these individuals remain heavily involved in the field. A listing of the major contributors to the field of clinical neuropsychology is beyond the scope of this chapter and involves a wide variety of individuals from different theoretical backgrounds and disciplines. In terms of major contributors to the development of neuropsychological assessment, such a list would have to include the following: Benton, Butters, Goldstein, Goodglass, Halstead, Kaplan, Meier, Parsons, Reitan, and Spreen. Citations and representative examples of their work can be found in various textbooks and merit review by those interested in obtaining more information

about clinical neuropsychology (e.g., see Grant & Adams, 1996; Lezak, 1995; Naugle, Cullum, & Bigler, 1998; Spreen & Strauss, 1998).

The concept of “approaches” to neuropsychological assessment can be considered from several perspectives. Historically, the “fixed” vs. “flexible” test battery approach was a topic of discussion and even heated debate for many years. Since many contemporary neuropsychologists tend to utilize a “core” battery of favorite measures for many of the patients they examine, but modify the assessment by adding or deleting specific tests depending upon the clinical population and/or referral issues in question, this has become an essentially moot issue. Furthermore, the modification of even the most “fixed” of test batteries to address individual patient needs and referral questions is now commonplace. Thus, the distinction between fixed and flexible approaches to neuropsychological assessment is somewhat arbitrary, although some discussion of the central issues involved in test selection and the composition of “standard” or core test batteries is in order.

4.11.2.1 The Standard Battery

In its strictest sense, a fixed standard battery approach involves the administration of the same set of tests to all patients, regardless of

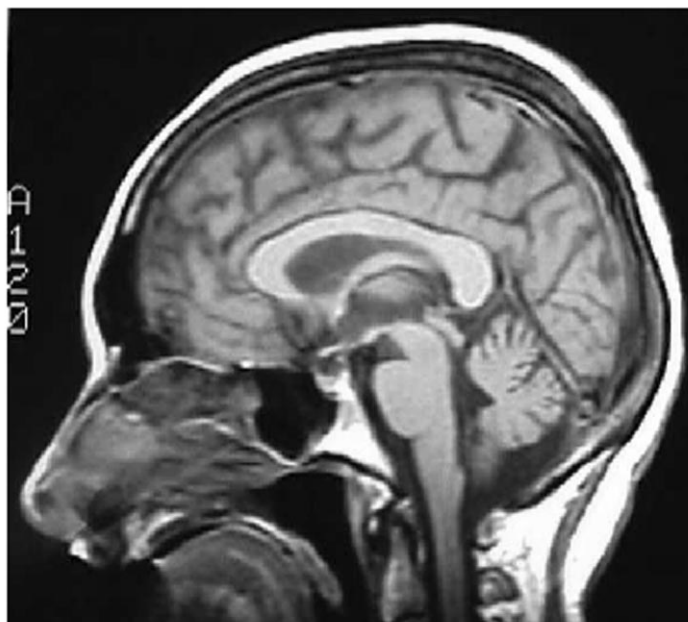


Figure 2 MRI depicting normal gross brain structure in a patient with severe dementia.

diagnostic or referral question, level of impairment, patient complaints, or clinical presentation. Applied too rigidly, such an approach may not yield appropriately detailed information regarding a specific area of deficiency (e.g., memory) unless that ability area is adequately represented in the standard or core battery. The application of an extensive omnibus test battery to all individual cases may also prove less efficient in some instances by oversampling behavior. For example, in the case of a rigid standard battery, a lack of impairment on a higher level ability measure may be followed by the routine administration of a similar task that tends to be less sensitive to deficits in that same domain. However, it has also been argued that a standard battery approach may allow for a broader assessment of neurobehavioral abilities in some settings, and hence, may detect cognitive impairment that a less comprehensive battery might miss. For example, a patient referred for memory assessment might be administered only a test of global intelligence (e.g., Wechsler Adult Intelligence Scale-Revised; WAIS-R) and selected measures of memory (e.g., the Logical Memory and Visual Reproduction subtests from one of the versions of the Wechsler Memory Scale are commonly used in various settings). Such an evaluation may not be sensitive to deficits in executive functioning or higher-order cognitive integration that might be manifest upon more detailed assessment and that can impact memory test performance. It should be noted, however, that the limited comprehensiveness of the battery would be a more

important issue than whether this represents a fixed or more flexible assessment approach.

Perhaps the best example of a standard neuropsychological battery is the group of tests that was originally created in the 1940s by Ward Halstead and further developed by one of his most prominent students and colleagues, Ralph Reitan. Halstead had strong interests in the assessment of biological intelligence and in the quantification of brain-behavior relationships, particularly at a time when localization of cerebral lesions was often a focus of neuropsychological evaluations. The work of Halstead and Reitan contributed monumentally to the development of American neuropsychology, and the Halstead-Reitan Neuropsychological Battery (HRB) continues to be used in many settings, whether used *in toto* or by its component tests.

The core HRB consists of the following tests and the primary corresponding cognitive functions (also see Dodrill, 1997): Category Test (abstract reasoning, logical analysis), Tactual Performance Test (complex psychomotor problem solving); Trail Making Test (psychomotor speed and cognitive sequencing); Aphasia Screening Test (brief assessment of language and graphomotor construction); Speech Sounds Perception Test (verbal/auditory attention); Seashore Rhythm Test (nonverbal auditory attention); Finger Tapping Test (fine motor speed); Grip Strength (gross motor strength); Sensory-Perceptual Examination (basic sensory and perceptual abilities); and Tactile Form Recognition (sensory perception, agnosia). Composite summary indices from the core

HRB can also be calculated in order to provide an overall index of cognitive impairment. These include the General Neuropsychological Deficit Scale (GNDS), which is derived from 42 scores from the HRB (0 = no impairment to a maximum possible score of 168; see Reitan & Wolfson, 1993), the Average Impairment Rating (Russell, Neuringer, & Goldstein, 1970) which ranges from 0.00 (no impairment) to 5.00, and the original Halstead Impairment Index (HII; Halstead, 1947), which utilizes a proportion of HRB test results based upon cutoff scores from selected tests (0.0 = no impairment to 1.0, reflecting impairment on all of the key variables). Like IQ scores, such summary indices suffer from a number of limitations (e.g., see Lezak, 1995, pp. 23–24), although they can arguably provide some useful reference information regarding an individual's overall functioning if not overinterpreted or used too rigidly.

Reitan and Wolfson (1996) provide a detailed overview of the HRB in terms of its application and interpretation, and many of the assessment and interpretation issues they discuss have relevance beyond the HRB *per se*. For example, the authors include a discussion of four of the principal ways in which test data can be interpreted: (i) level of performance; (ii) pathognomonic signs (i.e., specific test findings/deficits that strongly suggest brain dysfunction and are uncommon in nonbrain injured subjects); (iii) patterns of test results (e.g., those typifying different disorders); and (iv) comparison of test results using each side of the body (e.g., right vs. left hand) to examine hemispheric lateralization issues.

Clinicians vary with respect to their relative reliance upon each of these interpretive strategies, and depending upon the case, some findings may need to be given greater weight than others. For example, the pathognomonic sign of focal motor weakness in the absence of a peripheral motor injury clearly merits careful attention to the possibility of contralateral cerebral dysfunction. Similarly, a visual field cut (e.g., hemianopsia) should alert the clinician to focal cerebral involvement in the appropriate corresponding neuroanatomical area.

In other cases, the pattern of test results can be more important than the level of performance. Test findings that fall technically within normal limits may show a pattern that is suggestive of cerebral dysfunction. To illustrate this point using a more familiar measure, a patient may demonstrate overall scores within normal limits on the WAIS-III or WAIS-R, yet show relative weaknesses on those subtests most sensitive to cerebral dysfunction which may reflect neurologically-based deficits. Similarly, a normal summary score on other neuropsychological

tests can occur even when subscores and/or qualitative performance characteristics suggest clear abnormality or demonstrate a pattern consistent with a particular disorder.

In practice, many clinicians who utilize the HRB administer a modified version of the battery that selectively includes individual HRB tests. For example, many settings that do not routinely use the core HRB nevertheless administer the Category Test as a measure of higher-order problem-solving and because of its high sensitivity to cerebral dysfunction. Alternatively, clinical evaluations relying upon the complete HRB often include additional measures to provide a more detailed assessment of areas of cerebral functioning not well represented in the core battery. For example, one area that often is supplemented is memory, since the core HRB does not provide much in the way of memory assessment, even though this represents a common complaint among neurological populations.

Luria's Neuropsychological investigation (Christensen, 1984) represents a battery of measures that involves the systematic application of the assessment techniques originally described by Luria (1966). The battery is organized into various sections assessing an array of abilities, including basic and higher-level motor functions, receptive and expressive language, memory, and higher-order thinking. Even though it is presented here as a "standard" battery, it involves a sequential hypothesis-testing approach that includes a variety of individual tasks and relies upon qualitative judgments. Many clinicians utilize some of the component tasks selectively in their examinations. Some of the more commonly used tasks are the "go-no-go" paradigm (e.g., "When I hold up one finger, you hold up two," and vice-versa) and what has come to be called the "Luria 3-step" motor sequence by some (i.e., rapidly alternating the hand from flat to fist to edge) in order to assist in the examination of aspects of executive functioning and cognitive initiation/inhibition.

A very different type of standard test battery involves computer administration. Although several such test batteries exist, one example is the Microcog: Assessment of Cognitive Functioning (Powell et al., 1993). This is a computer-administered and scored battery of tests designed to screen for gross cognitive impairment in adults. It is available in short and long forms consisting of 12 and 18 subtests that assess the following general areas: attention/mental control, memory, reasoning/calculation, spatial processing, and reaction time. The test software calculates standard scores and percentile comparisons using age-referenced norms

that range from 18 to 89. While Microcog may be useful in some screening situations, the limitations of the brief and selective nature of the tasks, the lack of alternate forms, and the limitations inherent in computer-based assessment must be carefully considered in light of the specific clinical questions and patient populations in need of evaluation.

4.11.2.2 Test Batteries for Specific Populations

Various standard test batteries have been developed and/or assembled for specific purposes and populations, and many of these are reviewed in the excellent neuropsychological test compendiums by Spreen and Strauss (1998) and Lezak (1995). A few of the more commonly used batteries include the following:

(i) The consortium to establish a registry for Alzheimer's disease (CERAD) neuropsychological assessment battery (Morris, Heyman, Mohs, et al., 1989) was developed for use in patients with known or suspected dementia. This is a 30–45 minute examination of cognitive skills that includes measures of orientation, verbal fluency, naming, verbal learning and memory, and graphomotor constructional skills. Good reliability and validity have been demonstrated, and normative data for older adult populations are available (Welsh et al., 1994).

(ii) The NIMH core neuropsychological battery (Butters et al., 1990) was developed to evaluate cognitive changes associated with HIV infection. Both a brief (1–2 hour) and an extended (7–9 hours) battery comprised of standard clinical neuropsychological tests and measures used in cognitive psychology were assembled in order to assess the following cognitive ability areas: premorbid IQ, attention, speed of information processing, learning and memory, abstract reasoning, language, visuo-perceptual and graphomotor constructional abilities, and psychomotor skills. It includes tasks that provide information regarding patterns of cognitive deficits as commonly seen in degenerative conditions that affect cortical and particularly subcortical cerebral functions.

(iii) The Pittsburgh Occupational Exposures Test battery (Ryan, Morrow, Bromet, & Parkinson, 1989) was developed to evaluate those cognitive functions most commonly affected by exposure to environmental toxins. It comprises a series of 15 standard and experimental neuropsychological measures that assess a variety of abilities in a reasonably brief amount of time (i.e., approximately 90 minutes). Factor analysis of the battery revealed the following five domains: general intelligence, attention, learning and memory, visuospatial, and psychomotor speed and manual dexterity.

As noted, many neuropsychologists develop their own groups of tests for selected populations, such that a core battery for dementia might have some overlap with a battery for epilepsy (e.g., perhaps in terms of intellectual assessment). Such batteries would likely have different components and stepdown procedures, however. For example, additional or more comprehensive memory measures might be selected for patients with epilepsy when lateralized temporal lobe dysfunction is known or suspected. Furthermore, an assessment of academic achievement skills would typically be more important in the young adult with epilepsy than in the older individual with Alzheimer's disease. Thus, the issue of "core" test batteries depends largely upon neuropsychological training and experience, and many clinicians will use somewhat different sets of tests for different populations. In any case, it is important for the neuropsychologist to be familiar with a wide array of neuropsychological measures, not only in order to stay current with new test developments and research, but also to have a large test repertoire from which to select measures for individual or unique cases, as well as to be best prepared to review test results from other centers.

4.11.2.3 The Hypothesis-driven Approach

The hypothesis-driven or more flexible approach to neuropsychological assessment espouses the selection of tests based on referral questions, known or suspected diagnoses/pathology, and individual patient complaints and symptom presentation. Such examinations tend to be more "customized" or individually tailored, even though a relatively standard "core" battery of tests may be included as part of the evaluation. Tests may also be added or removed from the planned evaluation as the testing process progresses, depending upon performance results and patterns of strengths and weaknesses observed on various measures. As noted, many neuropsychologists administer a core group of tests to many or most of the patients they examine, subsequently tailoring aspects of the evaluation to follow up on areas of particular clinical interest. For example, a patient referred for evaluation of memory complaints might undergo a standard battery of measures that provides an evaluation of multiple cognitive domains, although the focus of the examination might be a more extensive assessment of various aspects of memory designed to thoroughly examine the patient's complaints or known/suspected pathology. Care must obviously be used in test selection, lest an examination become too cursory or

overly focused, or, at the other extreme, too lengthy and redundant. For example, an examination that consisted only of memory measures to evaluate a patient's memory complaints would neglect other areas of functioning (e.g., attention/concentration, executive functioning, language disturbance) that might play a prominent role in an individual's complaints of memory problems. Another issue is that once an area of deficiency is identified, the question arises as to how much exploration/assessment of the deficit is necessary. Certainly addressing the referral question is of great importance, and to the extent that this has been addressed, the clinician may decide to stop the evaluation at that point. In other cases, however, such as in the differential diagnosis of Parkinson's disease-related cognitive decline vs. depression vs. early Alzheimer's disease, more detailed exploration of multifaceted aspects of memory and other neurocognitive functioning would be required beyond the question, "Is there evidence of cognitive impairment?" that is often posed in certain settings.

4.11.2.4 Quantitative and Qualitative Examinations of Neurobehavioral Competence

Quantitative assessment refers to the use of test scores for interpretation, and qualitative examination often refers more to observations that pertain to the process by which patients perform tasks. To illustrate this latter point, even though a particular score on a test reflects a certain level of proficiency, that level of achievement may be attained via different cognitive strategies. An inefficient trial-and-error approach to a problem-solving task, for example, may yield the same final result as a more systematized strategy, even though these approaches reflect different underlying processes. Furthermore, the reliance upon either of these processes may have very different implications for patients in terms of their effectiveness in everyday functioning and in their ability to cope with novel situations. This approach to neuropsychological assessment is perhaps nowhere better illustrated than in the Boston Process Approach to clinical neuropsychology (see Milberg, Hebben, & Kaplan, 1996, for an overview) which incorporates an examination of the process or manner in which tasks are performed rather than relying upon final or summary scores alone. In some approaches to neuropsychological assessment, there is an emphasis on specific test scores, with less attention paid to how scores are achieved, the cognitive functions underlying those scores, the relative preservation of component functions

that may be indicated, and the manner in which those scores may relate to the patient's functioning in everyday life situations.

As noted, many clinical neuropsychologists implement aspects of flexible and standard, quantitative and qualitative approaches to assessment in practice, and these should not be viewed as mutually exclusive. Perhaps more important is the approach to test interpretation and performance analysis, which represents a multidimensional process. As noted earlier, the dimensions of level and pattern of performance on neuropsychological tests are very important to consider. In some cases, the level of performance may suggest normal functioning, while the pattern of test results suggests the presence of an abnormal process. Along these lines, it is important to keep in mind that scores in the "normal" or "average" range may reflect impairment in some individuals compared to their premorbid or baseline level of functioning. For example, the average IQ and memory scores in the previously high-functioning college professor who sustained a traumatic brain injury (TBI) may represent a significant decline in functioning, even though the obtained scores remain within the "normal" range (e.g., see Naugle, Cullum, & Bigler, 1990). Furthermore, if the patient achieves a good score on a particular test, but does so in an abnormal manner, this, too, may reflect altered mental function. Careful consideration of these factors, along with the pattern of test results and the degree to which the findings are consistent with deficits commonly seen in TBI, are essential in arriving at a correct diagnosis. Furthermore, the quantitative and qualitative information gained through a comprehensive neuropsychological examination may be very important in providing appropriate feedback and making realistic recommendations to the patient. Thus, as with the combination of standard and flexible test batteries, utilization of quantitative as well as qualitative data from the neuropsychological evaluation can be of extreme importance.

4.11.2.5 Other Issues in Test Interpretation

Regardless of a clinician's primary approach to neuropsychological assessment, the determination of what tests are to be given to each patient requires forethought and planning, and may vary depending upon the patient's age, level of education, socioeconomic status and ethnic background, and clinical presentation. For example, the neuropsychological evaluation of a retired physician following a mild brain injury would require the administration of a different set of measures than an 87-year-old laborer referred for diagnostic confirmation of

Alzheimer's disease who is known to have at least a 10-year history of progressive decline. As a more extreme example, the workup of an elderly individual with three years of education and limited English speaking abilities would require careful test selection to avoid measures with high cultural and educational biases.

4.11.2.5.1 Test cutoff scores

The use of "cutoff" or "cut" scores is another issue that merits comment, because this comes up frequently in clinical practice and often serves as a topic of debate. The cutoff score is based on the notion that most normal individuals tend to perform above a certain level on a given test, while most brain-injured individuals score below that level. From a statistical perspective, the notion of cutoff scores holds some merit, since many test results are normally distributed, and depending upon sensitivity and specificity qualities desired, cut points in a distribution can be set in order to maximize one or the other or both. This approach will produce definable true-positive and false-positive hit rates for groups, although statistical inferences may differ when applied to any given individual case. As a result, definitions of what constitutes "impairment" vs. "normalcy" may vary depending upon a variety of factors such as age and education, to take two of the more extensively studied demographic variables. Different cutoff scores for impairment would be needed to adjust for the effects of age on a given test, since age has a significant effect on so many cognitive measures. Using the original cutoff score of 90 seconds for the Trail Making Test—Part B, for example, would result in many healthy elderly individuals being misclassified as "impaired," when in fact their performances on this test may fall well within normal limits or even above average for their age. Similar considerations must be made with respect to different educational levels and estimated premorbid intellectual functioning, which can have a profound impact on the interpretation of certain neuropsychological test scores.

4.11.2.5.2 T-scores

Heaton, Grant, and Matthews (1991) provide normative reference values (*t* scores) for an extensive battery of tests that includes the HRB. This work, and the accompanying supplement for the WAIS-R (Heaton, 1992), represents a monumental contribution to the field by providing age-, education-, and where appropriate, gender-adjusted standard scores for an array of neuropsychological tests. The use of demographically-corrected *t* scores (derived from the same normative population) places

all test results on the same metric (mean = 50, SD = 10, with lower scores reflecting poorer performances), thereby allowing for ready comparisons across areas of cognitive functioning. For example, questions such as, "Is memory more impaired than expected given an individual's background and overall level of functioning?" can be addressed through the use of such standardized scores. Also, the effects of age and education on specific test results can be more readily appreciated, even when such factors may be thought to have little influence on a given task. For example, if it takes a 30 year old female with a high school education 90 seconds to complete Trails B, the corresponding *t* score is 38, which reflects a mild impairment, just over one standard deviation below average. If that individual were 75 years old, however, that same level of performance would fall in the above-average range (*t* score = 56).

The use of *t* scores allows for an individual's performance on specific tests to be compared with results from healthy groups of subjects of similar gender, age, and educational backgrounds rather than relying upon strict cutoff scores in order to help determine the degree of normality–abnormality of findings. If enough scores fall below expectation and/or form a consistent pattern of deficits, the likelihood of cerebral dysfunction is increased.

Given the error variance of any behavioral measure, *t* scores, like IQ or other standard scores, should not be used as rigid neurobehavioral markers that are absolute or "true" scores, but rather as interpretive guidelines to assist in evaluating levels and patterns of performance in a particular case. Furthermore, caution must be used in the strict application of interpretive scores and guidelines derived from normal populations to cases of brain-injured individuals (e.g., see Reitan & Wolfson, 1996). Another potential risk of over-reliance upon standardized scores in neuropsychological interpretation is that it can potentially give the (particularly inexperienced) clinician a false sense of security, when in fact, it is the neuropsychologist's training, knowledge base, experience, and skill in interpreting various aspects of test performance and behavior that result in valid neuropsychological conclusions regarding cerebral integrity.

4.11.2.6 Computer Interpretation of Neuropsychological Assessment Results

Computer-derived interpretive programs have been developed for several standard test batteries (e.g., HRB, Luria–Nebraska Neuropsychological Battery). While such programs

attempt to make general statements regarding the likelihood and even pattern of cerebral dysfunction based upon normative and neuropathological data and “typical” profiles, these programs may lend a false sense of interpretive security to those with more limited training and experience in clinical neuropsychology. While certain features of some such programs are arguably useful in certain situations (e.g., rapid scoring and normative referencing of results), they must be used with caution. Because brain damage affects individuals in different ways (i.e., a lesion in a specific location may produce different symptoms across individual patients depending upon a host of interindividual neuroanatomic, neuropathologic, genetic, experiential, and personality factors), any blanket statements about the nature of neurobehavioral disturbance in an individual must be examined carefully within the context of the patient in question. Particular care must be used when computer interpretations yield purported localization indices, since this process requires knowledge of underlying neuroanatomical systems, neuropsychiatric disorders, and individual neurobehavioral variations in order to optimize diagnostic accuracy. Attempts to distill numerically the complex process of clinically weighing various test scores, combinations of results, and qualitative performance features into “focal” interpretive patterns is a most challenging prospect, indeed, and erroneous interpretations can be rendered through over-reliance upon some of the available test interpretation programs.

4.11.2.7 Cognitive Screening

In keeping with the current managed-care *Zeitgeist*, neuropsychological evaluations should be appropriately detailed and comprehensive, while at the same time designed to provide the maximum amount of information in a time- and cost-efficient manner. It should be kept in mind, however, that brief cognitive assessments should not be carried out at the expense of thoroughness. In some cases, a brief examination may be adequate to address referral questions, as in the case of documenting cognitive impairment in a patient with severe dementia. A thorough examination of such a patient might be accomplished readily through the use of cognitive screening tools, and in fact, the administration of an eight- or 10-hour battery of measures in such a case might not yield any more diagnostic or clinically useful information than a brief assessment. However, evaluations that are overly brief may be insensitive to mild or even more significant cognitive deficits. Take, for example, the ever-

popular MMSE, which is arguably the most widely used cognitive screening tool and provides a very brief assessment of orientation, simple language, and recent memory skills. Whereas such instruments have utility in quantifying gross level of impairment when more severe brain dysfunction exists, scores that fall in the “normal” range on this test do not rule out cognitive abnormality. That is, simple screening measures such as the MMSE tend to have a high false-negative rate (i.e., identifying patients as being intact, when in fact, they are not). For example, diagnostic error rates using traditional cutoff scores on the MMSE (i.e., <24/30) can be around 15% even among patients with a diagnosis of possible or probable Alzheimer’s disease (Cullum & Rosenberg, 1998); in less impaired populations, the false-negative rate likely will be higher. Although the use of cognitive screening tools like the MMSE is preferable to a more idiosyncratic and non-standardized clinical mental status evaluation, the limitations of any brief cognitive screening tool must be kept carefully in mind. Along these lines, as with other measures of cognitive function, the use of cutoff or specific reference scores on such tests must be carefully considered in light of the individual patient’s background. Demographic factors often influence cognitive tests, and appropriate norms should be used even in the case of brief screening tools (e.g., see Crum, Anthony, Bassett, & Folstein, 1993 for age- and education-corrected MMSE scores).

4.11.3 METHODS OF NEUROPSYCHOLOGY

Neuropsychological assessment provides the most sensitive and comprehensive means by which brain function and cognition can be assessed given our current state of technology. The techniques are used widely in various settings across many patient populations, and are recognized by the American Academy of Neurology as accepted diagnostic procedures (Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology, 1996). Neuropsychologists often serve as consultants to neurologists, neurosurgeons, psychiatrists, clinical psychologists, and other medical and health professionals.

Neuropsychological tests are the means by which neurobehavioral samples are elicited, and specific test scores represent summary statements about observed behaviors and cognitive capabilities. Obtaining valid behavioral samples is critical, as is a thorough knowledge of the range of responses made by patients with and without various neurobehavioral disorders. As

noted in neuropsychology training programs, neuropsychological tests are no better than the clinician interpreting them. Thus, the best tests in poor or inadequately trained hands may fail to yield the correct neurobehavioral conclusions, and it is the clinician that is the instrument of evaluation—not the tests alone. Appropriate specialized training in neuropsychology is essential to conduct detailed neuropsychological evaluations beyond a cognitive screening level, and this is the topic of a section later in this chapter.

4.11.3.1 Clinical Interview and Background Information

Prior to a discussion of individual cognitive domains and specific representative measures, an overview of some of the other information that is essential to the neuropsychological examination is in order. In addition to a formal examination of mental functions, information pertaining to the patient's history, presenting complaint, and overall functional status is essential. In order to interpret results from any neurobehavioral task, an understanding of the behavioral geography of the brain and the underlying cerebrocortical systems is critical. It should be kept in mind that many cognitive processes represent inter-related phenomena and that individual tasks seldom assess highly specific skills or involve circumscribed cerebral areas in isolation. Second, background variables of the patient, including their age, level of education, SES, and cultural background, must be kept in mind while asking them to perform various tasks. Third, it should be kept in mind that the clinician interpreting the test data is the key to a good neuropsychological evaluation. Even the most sensitive tests in the hands of the inadequately trained clinician may yield erroneous conclusions. This is particularly true in cases of subtle behavioral abnormalities that may only be elicited through the use of specific cognitive tasks or detected via the careful and experienced observation of characteristic symptoms or behavior patterns. Table 1 provides an overview of much of the information that should be collected as part of the clinical interview within the context of the typical neuropsychological evaluation. Such data form a basis for the selection of assessment procedures as well as interpretation and integration of findings.

After obtaining this information, the selection of neuropsychological tests can be done, with particular areas of interest being stressed depending upon the referral question, nature of the disorder(s) under consideration, and patient complaints and presentation. In many settings, neuropsychological tests are administered by

Table 1 Cognitive domains typically assessed by the comprehensive neuropsychological evaluation.

1	Global cognitive/intellectual functioning
2	Academic achievement
3	Executive functioning/problem-solving
4	Language
5	Visuospatial processing
6	Attention/concentration
7	Learning and memory
8	Psychomotor functions
9	Sensory perceptual abilities
10	Psychological functioning

highly trained psychometrists or technicians. These individuals are responsible for standardized test administration and scoring under the supervision of the clinical neuropsychologist. This is a recognized standard practice in the field that is endorsed and/or utilized by a majority of neuropsychologists (Sweet & Moberg, 1990).

4.11.3.2 Neuropsychological Measurement of Brain Function

The comprehensive neuropsychological evaluation includes the administration of tests that comprise multiple cognitive domains. Although the specific tests selected by individual neuropsychologists may vary, the domains listed in Table 2 are commonly represented.

Excellent descriptions and reviews of the most frequently used cognitive measures (as well as normative data for some tests) can be found in Lezak (1995) and Spreen and Strauss (1998), and the reader is encouraged to consult these comprehensive reference sources. Table 3 presents some of the more commonly used tests to assess these cognitive domains. It should also be noted that there may be significant overlap between some of the listed tests, and across domains. The tests are listed for illustrative purposes and are not intended to provide an exhaustive list or to reflect a prototypical clinical neuropsychological test battery.

The question sometimes arises as to why a neuropsychological evaluation cannot be extremely focused, that is, "Why not administer a single test of memory to a patient if that is the only presenting complaint?" The answer obviously is that by using only one test or assessing only one cognitive domain, the clinician might completely overlook the actual cause of the patient's reported difficulties. For example, an individual presenting for memory disturbance may actually have primary difficulties in attention and concentration which preclude their ability to efficiently encode new material. When that individual attempts to recall information at

Table 2 Common neuropsychological measures.

<i>Representative measures</i>	<i>Functions assessed</i>
<i>Intellectual</i>	
Wechsler Adult Intelligence Scale-3 (WAIS-III) and Wechsler Adult Intelligence Scale-Revised (WAIS-R)	Global cognitive capacity/ Intelligence
National Adult Reading Test-Revised (NART-R)	Premorbid IQ estimation
<i>Global cognitive</i>	
Dementia Rating Scale (DRS)	Presence and level of dementia
Mini-Mental State Examination (MMSE)	Gross cognitive screening
<i>Academic achievement</i>	
Wide Range Achievement Test-3 (WRAT-3)	Reading, spelling, math
Pearowstart Individual Achievement Test-Revised (PIAT-R)	Reading, spelling, math
Wechsler Individual Achievement Test (WIAT)	Basic academic skills
<i>Executive function and problem-solving</i>	
Wisconsin Card Sorting Test (WCST)	Cognitive flexibility/problem-solving
Category Test	Abstraction and reasoning
California Sorting Test (CST)	Cognitive flexibility/idea generation
Similarities (WAIS-III, WAIS-R)	Abstraction
Trail Making Test (Part B)	Mental sequencing/flexibility
Raven's Matrices (standard, colored, advanced)	Nonverbal reasoning
<i>Language</i>	
Vocabulary (WAIS-R)	Word knowledge
Boston Naming Test (BNT)	Confrontation naming
Word fluency; Controlled Oral Word Association	Verbal fluency
Token Test	Comprehension
Boston Diagnostic Aphasia Examination (BDAE)	Overall language function
Aphasia Screening Test	Language screening
<i>Visuospatial</i>	
Block Design (WAIS-R)	Visuoconstructional ability
Clock drawings (command and copy)	Graphomotor construction
Cross drawings (command and copy)	Graphomotor construction
Rey-Osterrieth Complex Figure (copy)	Graphomotor construction/planning
Hooper Visual Organization Test (HVOT)	Visuoperceptual integration
Object Assembly (WAIS-R)	Visuospatial integration
Line bisection, Visual cancellation tests	Hemispacial inattention
Ruff Figural Fluency Test	Nonverbal fluency
<i>Attention/Concentration</i>	
Digit Span forward (WAIS)	Simple attention
Digit Vigilance (number cancellation)	Sustained concentration and attention
Paced Auditory Serial Addition Test (PASAT)	Auditory concent/processing speed
Digit Symbol (WAIS) or Symbol Digit Modalities Test	Visual attention/psychomotor speed
Trail Making Test (Part A)	Visual tracking/psychomotor speed
Continuous Performance Test (various versions)	Sustained vigilance and reaction time
Visual Sustained Attention Test (VSAT)	Sustained concentration and attention
<i>Learning and memory</i>	
Wechsler Memory Scale-3 (WMS-III) and Wechsler Memory Scale-Revised (WMS-R)	Global memory function
California Verbal Learning Test (CVLT)	Verbal learning and memory
Logical Memory (WMS-III, WMS-R)	Verbal memory
Visual Reproduction (WMS-III, WMS-R)	Nonverbal memory
Rey-Osterrieth Complex Figure (recall)	Complex nonverbal memory
Hopkins Verbal Learning Test	Brief measure of verbal memory
Benton Visual Retention Test	Nonverbal memory
Warrington Recognition Memory Test	Verbal/nonverbal recog. memory
Rey Auditory Verbal Learning Test	Verbal learning and memory
Buschke Selective Reminding Test	Verbal learning and memory
<i>Motor and sensory</i>	
Finger Tapping Test	Fine motor speed
Hand Dynamometer	Grip strength
Grooved Pegboard	Fine motor dexterity
Luria three-step hand sequence	Coordinated/integrated sequencing
Sensory-Perceptual Examination (Reitan-Klove)	Sensory and perceptual skills

Table 3 Information to be collected from the history and behavioral observations.*History of symptoms*

Description of present illness and specific symptoms (e.g., Why is patient being evaluated?)
 Nature of onset (e.g., insidious vs. acute; are symptoms new or a recurrence?)
 Duration
 Course of symptoms
 Associated behavioral or personality/emotional changes

Past medical history

Previous neurological disease
 Previous psychiatric disease (psychiatric treatment, known diagnoses, etc.)
 Significant head trauma (details: presence and length of LOC/coma; PTA—assess)
 Seizures (circumstances of onset, frequency, characterization of seizures)
 Other medical illnesses requiring hospitalization
 Toxic Exposure (may relate to work history)

Birth and developmental history

Birth complications
 Developmental delays (motor, language, intellectual)

Social history

Abnormalities in behavior past and present
 History of and current family and peer interactions

Educational history

Highest grade attained/ degrees awarded
 Specific problems learning
 Problems in school (e.g., truancy, being expelled, etc.)
 Typical grades in school (e.g., below average, average, above average)
 Best and worst subjects in school
 Reasons for termination of education (if relevant; also, if not HS, did get GED?)

Vocational history

Present occupational status (if retired, what was last position)
 Length of current job
 Nature of responsibilities
 Types of previous jobs
 Frequency of job changes
 Recent difficulties with work (self-report, work appraisals, raises, etc.)

Substance use history

Duration, frequency, and amounts of typical alcohol intake
 Use of other drugs
 Use of prescription medications (list current medications)

Family history

Neurologic disease in family (especially parents, siblings)
 Psychiatric disease in family
 Familial predilection for diseases (e.g., learning disabilities, dementia, epilepsy, etc.)
 Cause and age of death of parents and siblings

Psychiatric symptoms

Unusual or bizarre behavior
 Attention/concentration problems—distractibility
 Problems with social judgment
 Depression
 Anxiety
 Paranoid ideation/suspiciousness
 Hallucinations
 Delusions
 Tangentiality, looseness of associations

General appearance

Appropriateness of appearance for age
 Posture (examine for asymmetries)
 Facial expression
 Eye contact
 Personal cleanliness, grooming, dress

Mood

Range and appropriateness of affect, responsiveness to conversation/interview
 Depression
 Apathy
 Lability

Table 3 (continued)

Mania or hypomanic symptoms
<i>Behavioral</i>
Cooperativeness with examiner
Anxiety
Suspiciousness
Anger
Insight
Response to interview
<i>Motor</i>
General activity level
Abnormal posturing, movements, tremors, etc.
Gait (broad-based, shuffling, unsteady, symmetry of movement)
Handedness (and familial history thereof, particularly if left-handed)
<i>Speech</i>
Rate
Volume
Modulation
Expression
Dysarthria

a later time, they fail, thereby producing what subjectively seems like a deficit in memory. In actuality, their problem may not be at the retrieval or recall stage, but rather, at earlier stages of information processing. Similarly, a patient with aphasia may present with word-finding difficulty that might also mimic a memory impairment if the individual were not thoroughly evaluated. Some patients with frontal lobe damage demonstrate inefficient encoding secondary to organizational deficits, when in fact their ability to recall learned material may be normal.

The following section discusses the major cognitive domains typically assessed by the comprehensive neuropsychological evaluation, including an overview of some common representative measures from each domain. Each section begins with a clinical overview of the area to be examined and includes a brief discussion of relevant, less structured or “bedside” clinical examination/observation procedures prior to a presentation of specific psychometric assessment tools.

4.11.4 PRINCIPAL COGNITIVE DOMAINS FOR ASSESSMENT

4.11.4.1 Global Cognitive/Intellectual Functioning

The concept of intelligence has a long history and has been the topic of much discussion and controversy since the early 1900s. Various definitions of intelligence have been posited (Sternberg, 1997), and it is clear that intelligence is a multidimensional construct that is made up of many types of abilities that are demonstrated by each of us to varying degrees. Despite ongoing debate regarding the use of IQ measures,

the concept of global intellectual capacity has utility within the context of assessing cognitive functioning and adaptive ability as it relates to brain function.

From a clinical bedside perspective, the clinician must rely upon the patient’s educational and vocational history, in addition to their overall presentation style, use of language (when possible), and performance during various cognitive tasks to estimate a gross level of intellectual functioning. Next, one must begin to evaluate whether there seems to be evidence of gross intellectual decline based on their current presentation. Ideally, premorbid IQ and other cognitive test scores would be available for direct comparison, but this is rarely the case. Thus, detailed examination of this issue requires formal psychometric assessment, and based on specific cognitive test results and patterns of test scores, indices of current intellectual functioning as well as estimates of premorbid intelligence can be derived (Franzen, Burgess, & Smith-Seemiller, 1998).

The most commonly used intellectual assessment tools for adults are the Wechsler Intelligence Scales. In addition to the information provided by the various component subtests of the Wechsler Scales, some assessment of an individual’s global level of cognitive capacity is important in establishing an interpretive background for other specific test results. For example, knowing that a patient has a Full-Scale IQ (FSIQ) score in the borderline range can be very useful in accurately interpreting their low average to borderline performance on memory tests. Although some variability across cognitive skills is commonly seen within individuals, higher intellectual levels tend to be associated with more advanced skills in various domains.

Thus, the 75-year-old retiree with eight years of formal education who obtains an FSIQ of 120 would be expected to perform at a higher level on memory or attention/concentration measures than the elderly individual with a similar educational background, but an IQ in the low average range.

One of the limitations of intellectual assessment that is worth noting is the use of global IQ scores in depicting an individual's cognitive capabilities. An FSIQ of 100 can be arrived at in various ways, for example, by achieving an very high Verbal IQ and low Performance IQ score, or vice-versa. On the WAIS-R, note that each of the following sets of VIQ/PIQ scores result in an FSIQ of 100 for a 30-year-old: (i) VIQ = 121, PIQ = 79; (ii) VIQ = 80, PIQ = 138; (iii) VIQ = 100, PIQ = 101. Also, some subtest scaled scores may be very high and offset lower ones, resulting in an "average" overall IQ score. These examples represent significantly different cognitive processing styles and differential ability patterns, although these distinctions would be lost if there is an over-reliance upon global IQ scores.

From a neuropsychological perspective, intellectual tests such as the WAIS-R and WAIS-III can be quite useful in describing cognitive strengths and weaknesses and eliciting relevant neurobehavioral samples, even though these tests were not originally designed for assessment of cognitive dysfunction. However, many of the changes in the WAIS-R reflected in the WAIS-III, in fact, were influenced by neuropsychological principles, and the WAIS-III includes several new subtests that attempt to provide a more detailed exploration of cognitive abilities beyond "IQ" or "g" alone. Whereas intellectual tests tend to be less sensitive to cerebral dysfunction than many specialized neuropsychological measures of specific abilities, the relative sensitivity of some subtests to cerebral dysfunction have proven useful in detecting gross cognitive decline in some cases. For example, the Vocabulary subtest of the WAIS-R historically has been known as one of the tests that is least sensitive to acquired cerebral dysfunction. This subtest relies upon highly crystallized and overlearned knowledge, and the ability to define familiar words tends to be a function that is relatively less affected by acquired brain damage than many other tasks that require novel learning or reasoning. In contrast, the Digit Symbol subtest of the WAIS-R, which requires rapid psychomotor speed, new learning, and attention, is one of the more sensitive subtests to cerebral dysfunction. Knowledge regarding which subtests are more or less sensitive to brain injury, educational background, age, gender, and so on, is essential

in making inferences about cerebral dysfunction (see Heaton, Ryan, Grant, & Matthews, 1996). Pattern analysis of the Wechsler scales can thus be quite useful in the assessment of brain damage, as the subtests are differentially susceptible to the effects of acquired brain injury and can be further impacted depending upon the nature of the disorder and/or location of primary cerebrocortical involvement. These tests are described now in detail.

(i) *Wechsler scales.* (a) Wechsler Adult Intelligence Scale-Revised. The Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) is the most popular measure of intellectual functioning in adults, and has enjoyed widespread use since its introduction in 1981. A plethora of studies using the WAIS-R or its components has been published in the neuropsychological literature, and a great deal has been learned about its utility as a neuropsychological tool. This is underscored by the development and publication of the WAIS-R as a neuropsychological instrument (WAIS-R-NI; Kaplan, Fein, Morris, & Delis, 1991). This modification of the WAIS-R reflects the application of specific neuropsychological principles and procedures to the administration of the WAIS-R. For example, in individuals with difficulty responding to open-ended questions or those having initiation difficulties, the Information subtest of the WAIS-R-NI includes a provision for multiple choice responding. Some of the timed subtests include extended time limits for those with motoric slowing to permit a more valid assessment of visuoconstructional skills apart from the effects of motor dysfunction.

In settings where an overall estimate of intellectual functioning is desired, various short forms of the WAIS-R have been developed. One very brief example is the combination of Vocabulary and Block Design in the estimation of FSIQ (Brooker & Cyr, 1986; Silverstein, 1982). Such brief forms and lengthier procedures (e.g., Adams, Smigielski, & Jenkins, 1984) have shown reasonable correlations with the complete WAIS-R, and these estimates may be adequate in some situations. For example, in a case of head injury wherein the focus is upon more subtle difficulties with attention/concentration and memory, less of an emphasis may be placed on global intellectual status, particularly when there are no questions of a decline in this area. In selecting the shortened version of any test such as the WAIS-R, clinicians must rely upon experience with a large number of cases and with appropriately supportive literature. Among the individual subtests that show the strongest correlations with overall FSIQ scores is the Vocabulary subtest ($r = 0.81$), followed by Information ($r = 0.81$). At the other end of

the spectrum lies Digit Symbol, which, along with Object Assembly, shows a correlation of 0.57 with FSIQ. In general, the five Performance subtests, which rely more upon novel thinking and “fluid” cognitive skills, tend to be more sensitive to cerebral dysfunction than the more crystallized abilities tapped by the six Verbal subtests. Because the various subtests of the WAIS-R assess a variety of cognitive abilities, some of these will be discussed in more detail under specific cognitive domains.

(b) Wechsler Adult Intelligence Scale-III. The WAIS-III is a newly revised version of the Wechsler Adult Intelligence Scale (WAIS-III; Wechsler, 1997a). It contains a variety of changes and additional subtests, and some previously familiar measures have been dropped or are now optional. The addition of a matrix reasoning task and measures of sustained concentration and working memory promise to provide a more well-rounded evaluation of global cognitive functioning and may enhance the test’s utility in the context of the neuropsychological evaluation. Most of the familiar subtests remain, although these have been updated for the 1990s, with improved norms. Many subtests that include scenes for visual material now are presented in color, with larger figures and a more representative depiction of our multicultural society. As a quick review of the test evaluators and consultants will indicate, there is a decidedly more neuropsychological flair to the WAIS-III. As we have learned more about brain-behavior relationships and the effects of acquired cerebral dysfunction on intellectual functioning, test developers felt it important to incorporate aspects of this knowledge into the current test design. Table 4 lists the subtests for the WAIS-III.

As reported in the WAIS-III technical manual (Wechsler, 1997a), correlations involving VIQ, PIQ, and FSIQ scores between the WAIS-III and WAIS-R are high (0.94, 0.86, and 0.93, respectively). WAIS-III VIQ, PIQ, and FSIQ scores were found to be an average of 1.2, 4.8, and 2.9 points lower than the corresponding values on the WAIS-R. Obviously, given the recent release of this test, many studies of its neuropsychological utility and psychometric relationships to other neuropsychological measures will be forthcoming.

(c) The Wechsler scales and patterns in neurological populations: VIQ-PIQ Differences. Despite the utility of instruments like the Wechsler scales in neuropsychological evaluations, appropriate interpretation of scores and patterns of scores in cases of brain dysfunction requires extensive training in neuropsychology and familiarity with a broad range of neuropathological, neuropsychiatric, and normal

Table 4 WAIS-III subtests.

Vocabulary	Picture Completion
Similarities	Digit Symbol-Coding
Arithmetic	Block Design
Digit Span	Matrix Reasoning
Information	Picture Arrangement
Comprehension	Symbol Search
Letter-Number Sequencing	Object Assembly

conditions. For example, deficits on PIQ subtests are often discussed in association with right hemisphere dysfunction. Whereas this may hold true in many cases, decreased nonverbal test scores may occur for a variety of reasons (e.g., psychomotor slowing on timed tasks, greater sensitivity to acquired brain injury than most of the VIQ tasks). The following cases serve to illustrate this point.

First it should be noted that all three cases depict at least a 22-point difference between VIQ and PIQ scores, although there is obviously some variability across the profiles in terms of specific scaled scores. One initial impression might be that all three cases likely represent some sort of acquired brain injury, as 20-point differences between IQ indices are rare in the general population. A second level of interpretation, based in part on the clinical lore associated with the test, might include an inference about lateralization, that is, all three cases might be suggestive of greater impairment of right hemisphere functions, given the large VIQ-PIQ differences and relative deficits on the “nonverbal” subtests. As alluded to earlier, knowledge that a variety of factors may result in decreased PIQ scores can provide the clinician with information that is essential in neurodiagnostic situations.

Case 1 is an engineer in his 40s who sustained a right hemisphere stroke. CT findings (Figure 3) depict a hemorrhage deep in the right hemisphere that included intraventricular bleeding.

His WAIS-R profile of decreased nonverbal abilities and a 22-point VIQ-PIQ discrepancy is what is often associated with a “classic” pattern of right hemisphere dysfunction (see Figure 4). Whereas this holds true for this particular individual, Cases 2 and 3 illustrate that simplistic interpretations or overgeneralizations can lead to erroneous conclusions.

Case 2 is a 50-year-old nurse with a history of stroke four months prior to evaluation. MRI results (Figure 5) revealed a chronic egg-sized hemorrhage involving the left internal capsule and basal ganglia.

Her primary presenting symptoms included mild right-sided hemiparesis and mild difficulties with verbal fluency and word finding. Much of the decrease in PIQ subtest scores (Figure 6)

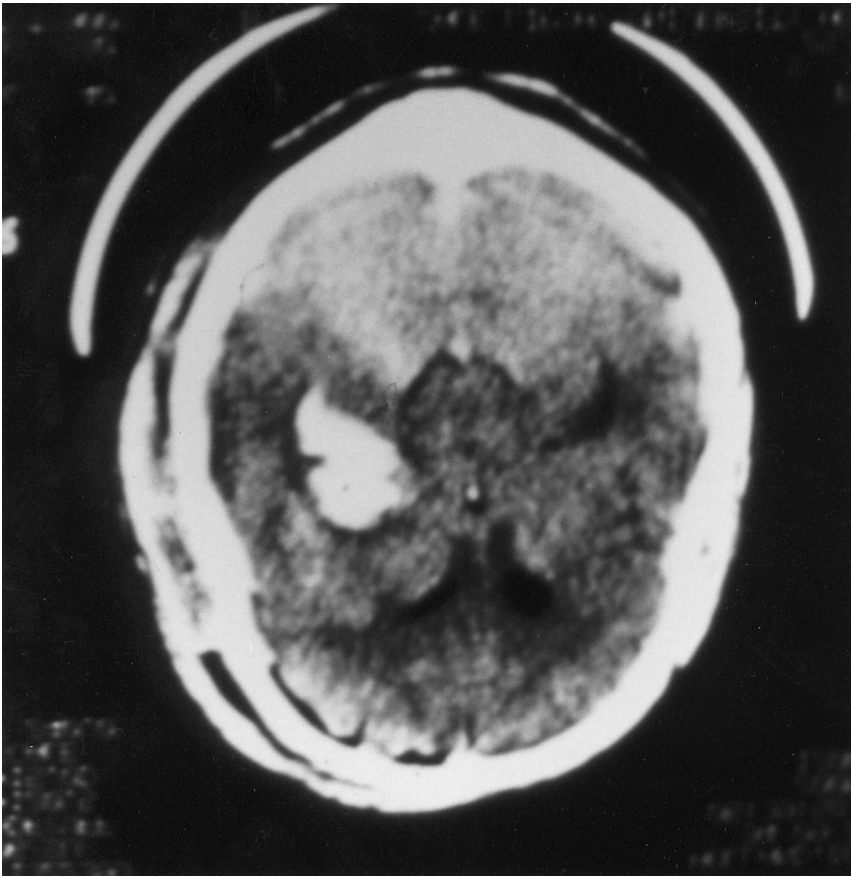


Figure 3 CT from Case #1 showing right hemisphere infarct.

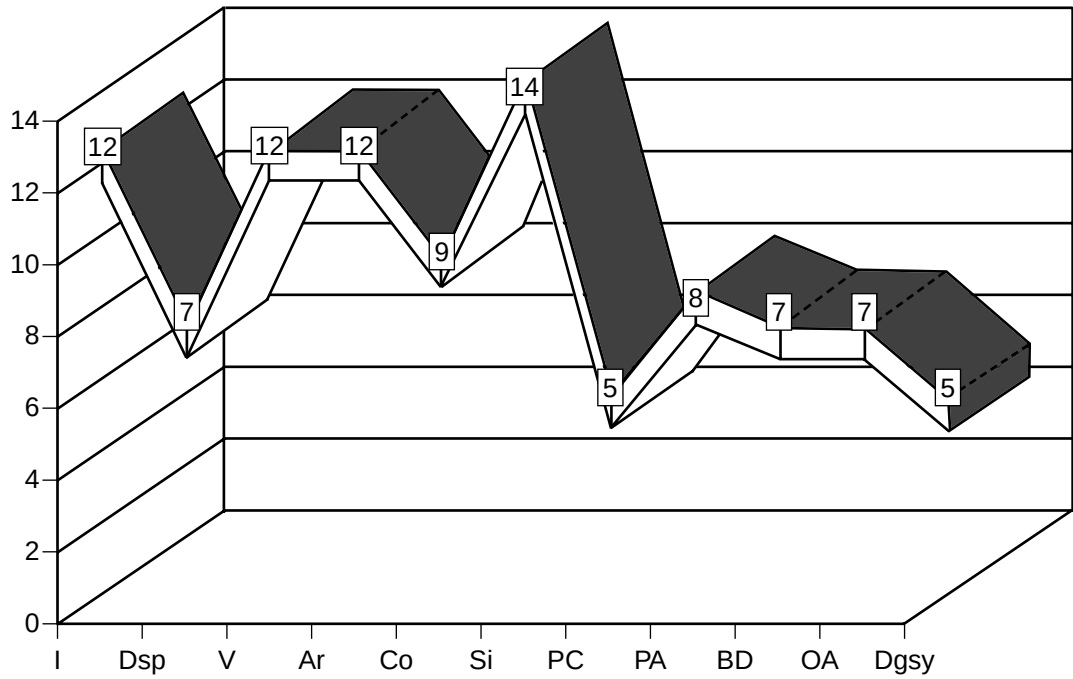


Figure 4 WAIS-R scaled scores from case #1. VIQ = 108, PIQ = 86, FSIQ = 98.

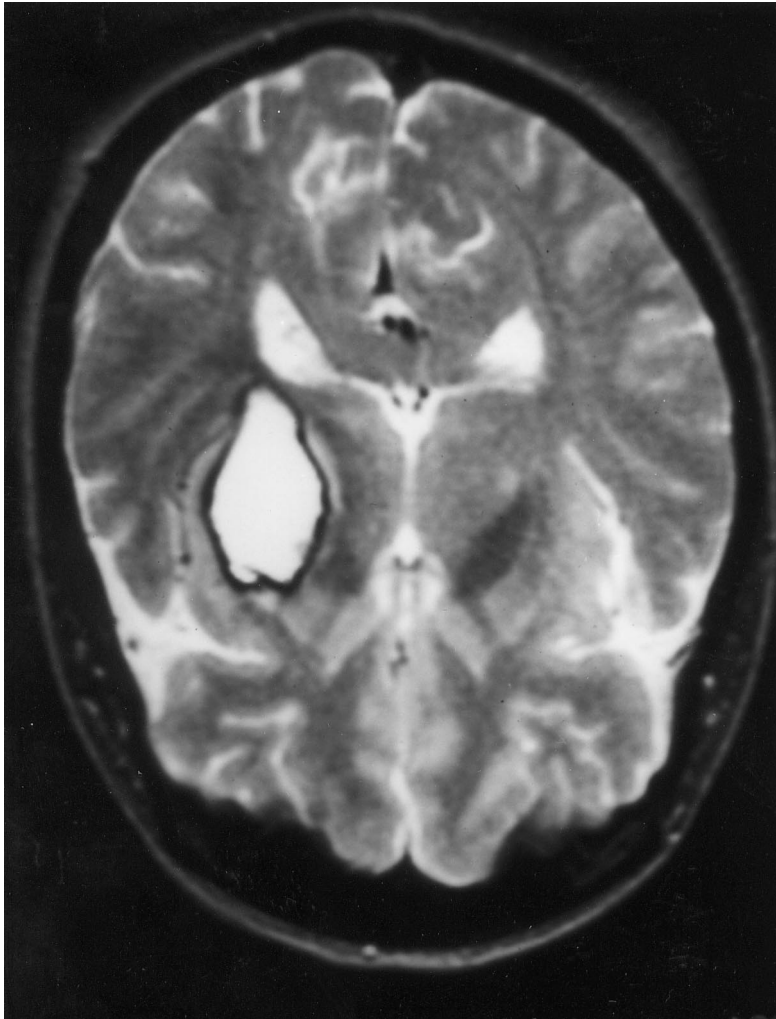


Figure 5 MRI showing left basal ganglia/internal capsule hemorrhage in case #2 (note that left is depicted on the right).

was felt to be related to the patient's hemiparesis and having to perform some tasks with her nondominant left hand. However, the relative sensitivity of nonverbal vs. verbal WAIS-R subtests to any type of acquired brain damage should be kept in mind, and in fact, additional time allowances on some of the PIQ subtests (e.g., Block Design) still resulted in evidence of cognitive impairment; thus, psychomotor factors alone could not completely explain the patient's difficulties in performing these tasks.

Case 3 results are from a 40-year-old long-term alcoholic who showed evidence of alcohol-related dementia. Axial MRI results (Figure 7) depict generalized sulcal widening and ventricular dilation.

The patient demonstrated generalized neuropsychological deficits, with particular difficulties in visuospatial abilities (see Figure 8). Results such as this are not uncommon in cases of extensive and long-term (i.e., 10+ years)

alcohol abuse, and this pattern of findings on the WAIS-R was amply supported by the patient's impaired performances on other visuoconstructional tasks.

(d) WAIS subtest "scatter." The issue of "scatter" among WAIS-R subtest scaled scores represents another area of clinical lore that merits comment within the context of assessing patients with known or suspected cerebral dysfunction. It has been observed by some that, in cases of brain damage, there tends to be a greater than average range of scores from highest to lowest across subtests. This represents a complex issue, since most individuals with or without brain damage have relative cognitive strengths and weaknesses, and as noted, the subtests differ with respect to their individual sensitivity to acquired cerebral dysfunction.

Finally, what constitutes clinically "significant" scatter across Wechsler subtest scores also

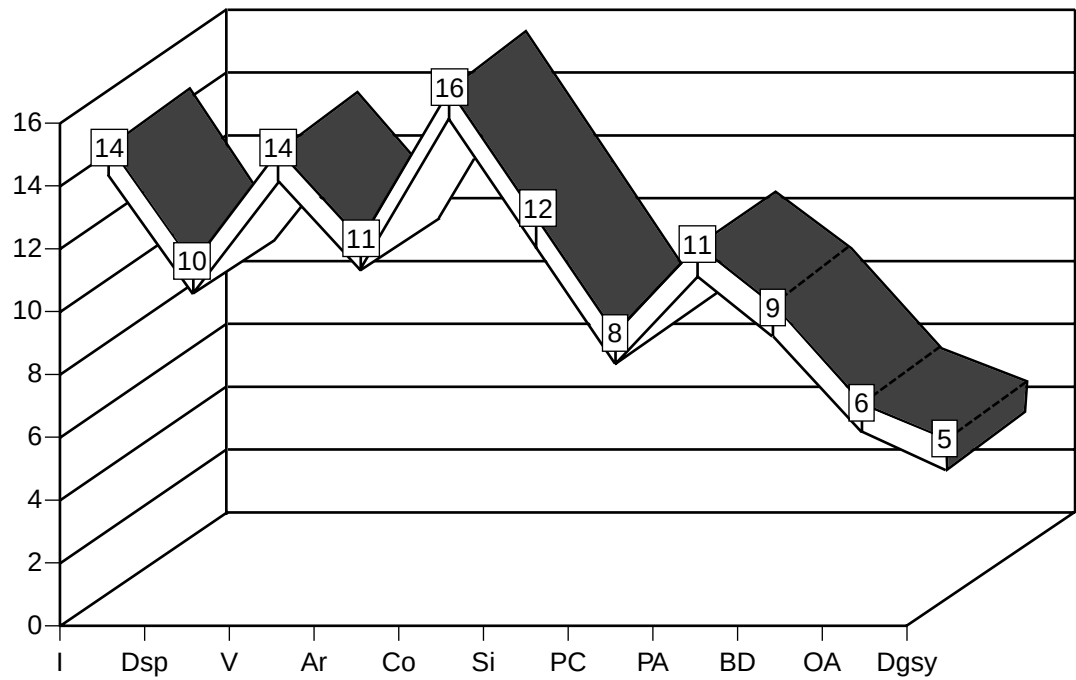


Figure 6 WAIS-R scaled scores from case #2. VIQ = 121, PQ = 101, FSIQ = 112.

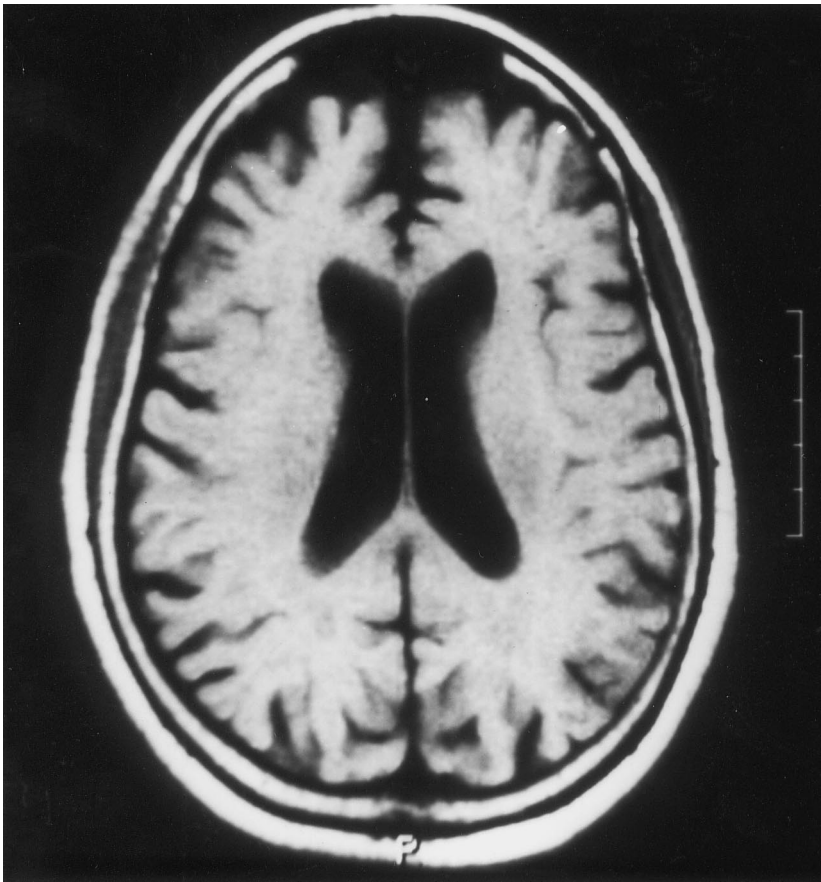


Figure 7 MRI results from case #3 showing generalized atrophy.

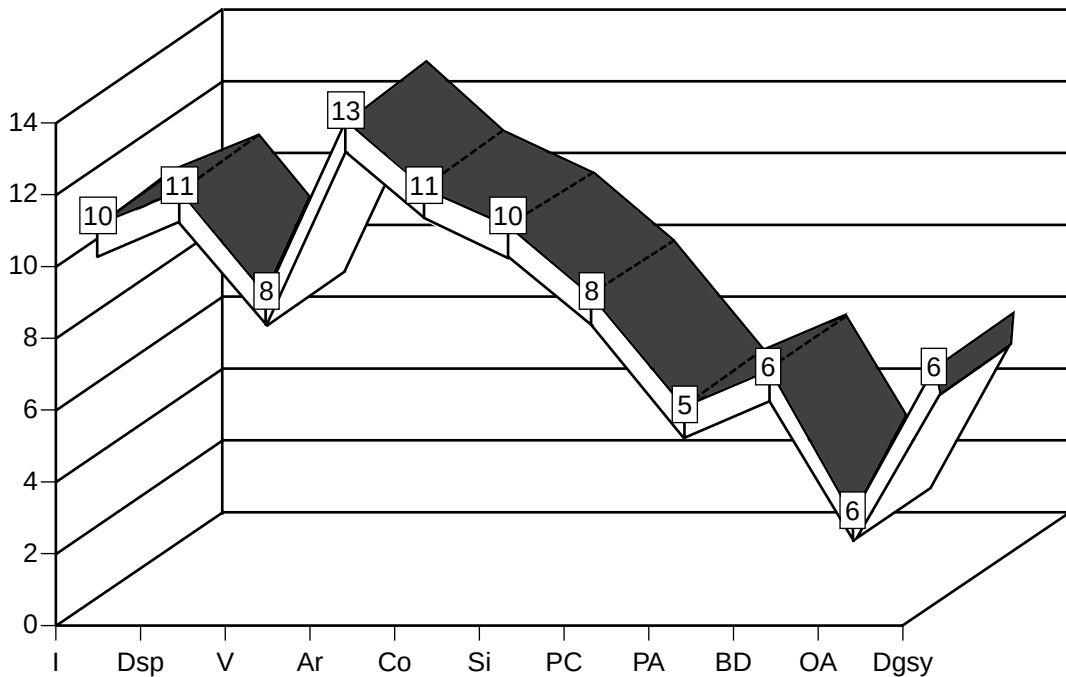


Figure 8 WAIS-R scaled scores from case #3. VIQ = 103, PQ = 78, FSIQ = 90.

merits careful review. In his APA Presidential address, Matarazzo (1990) provides data that clearly calls to question the notion of the relevance of “scatter.” For example, a table is presented (p. 1010) that was derived from previous work (Matarazzo & Prifitera, 1989) which includes the percentage of cases from the WAIS-R standardization sample that demonstrate various levels of scatter (the difference between highest and lowest of the 11 subtest scaled scores). While extreme score differences of greater than 10 points were rare (i.e., seen in less than 5% of the 1880 standardization sample subjects), discrepancies of 10 points were seen in approximately 9% of the sample, while differences of seven points occurred in almost half of the sample, and roughly 86% showed a difference of five points. Thus, it should be kept in mind that “scatter” among WAIS subtest scaled scores is commonly seen in non-neurological populations. With these numbers in mind, the issue of the neurological significance of scatter among Wechsler scaled scores should be considered cautiously.

(ii) *Ravens Progressive Matrices*. This series of tests includes the Colored (Raven, 1995), Standard (Raven, 1996), and Advanced (Raven, 1994) Progressive Matrices, providing different levels of assessment of nonverbal reasoning. In the Colored Matrices, subjects are shown a stimulus array at the top of a booklet, with one small piece cut away. Below this, six small pieces corresponding to the outline of the missing piece are presented;

however, each typically has a different pattern depicted. The task is to select the piece that best completes the stimulus array. The Colored matrices are particularly well suited for children and impaired elderly, and the Advanced matrices are most appropriate for high functioning individuals when an assessment of nonverbal reasoning is desired. IQ equivalent scores can be derived from the matrices, and their nonverbal nature (particularly the Colored version) allows for those with language impairments or limited fluency in English to apprehend the nature of the task quite readily in many cases.

(iii) *National Adult Reading Test-Revised (NART-R)*. This is a measure of estimated premorbid intellectual functioning (Blair & Spreen, 1989) as derived from the original NART (Nelson, 1982; Nelson & O’Connell, 1978). It comprises 60 printed words that are presented to subjects to read aloud. Thus, rather than relying upon any sort of integrative or associative abilities, the NART-R and its various derivations rely only upon sight-word reading. Most of the words included in the NART-R are phonetically irregular, and thus, correct pronunciation generally depends upon prior familiarity with the words. The stimuli are presented in order of increasing difficulty, and range from words like “asterisk” to “ennui.” As with any performance-based cognitive task, the NART-R does show some sensitivity to acquired brain dysfunction; nevertheless, it tends to remain relatively stable, even in the face of progressive dementia (Matt-Maddrey, Cullum,

Weiner, & Filley, 1996). The NART-R does have limitations in terms of its applicability to individuals with very high and very low intellectual status, but correlates well with Verbal and Full-Scale IQ in many cases when there is little or no intellectual decline.

(iv) *Dementia Rating Scale (DRS)*. This is a popular instrument for the assessment of global cognitive functioning in patients with known or suspected dementia (Mattis, 1988). It has been widely used clinically as well as for research purposes, and has proven to be a reliable and valid screening tool that typically requires only 20–30 minutes (up to 45 minutes in some demented individuals). The total number of possible points on the DRS is 144, and Table 5 provides a general description of level of impairment on this measure. It should be noted, obviously, that any such listing provides only a global guideline for interpretation, as results on the DRS do vary according to age, education, and other background factors.

The DRS is made up of the following subtests: attention, initiation/perseveration, conceptualization, construction, and memory. Because of the screening nature of the test, higher functioning or minimally impaired individuals may achieve normal scores yet demonstrate evidence of impairment on more sensitive tests. For example, the memory component of the DRS assesses general orientation and includes some simplistic recognition memory tasks, while only two tasks focus on the assessment of delayed recall.

Although designed as a tool to screen for the presence or absence of dementia, patterns on the DRS have also been shown to be useful in helping to distinguish some types of dementias, although it should be noted that once cognitive impairment becomes severe, such differentiation on the basis of cognitive test results can be difficult or impossible. Despite its relative brevity, total scores from the DRS are quite useful in documenting cognitive decline over time and have shown to be predictive of nursing home placement (Smith et al., 1994).

(v) *Mini-Mental State Examination*. This is a very popular screening measure of global cognitive status. It is particularly useful in cases of dementia and may be used even in severely impaired individuals (Folstein et al., 1975). It is

very brief (i.e., 5–10 minutes) and comprises 30 total points. Test items include an assessment of orientation to time and place, attention, language, constructional skill, and three word repetition and recall. It is recommended that standardized recall items (e.g., apple, table, penny or rose, ball, key) are used, as opposed to the original instructions of “name three objects” for patients to recall. Total MMSE scores correlate best with orientation and recall items, and significant variability can be seen among healthy subjects on three word recall (Cullum, Smernoff, & Thompson, 1993). Despite the limitations inherent in any brief measure of overall cognition, the MMSE has proven to be a valid and reliable screening tool, with utility in a number of populations and settings. Although the traditional cutoff for impairment is $< 24/30$, the MMSE is sensitive to age and education effects, and appropriate norms should be used (Crum et al., 1993; Tombaugh, McDowell, Krisjansson, & Hubley, 1996).

4.11.4.2 Academic Achievement

These tests sample a range of abilities related to academic or scholastic levels of attainment. A number of omnibus achievement batteries exist and each has its relative strengths and limitations depending upon the purpose of assessment. Within the context of the adult neuropsychological evaluation, various components of different achievement tests are often administered. The more common measures sample a variety of abilities related to academic achievement (e.g., spelling, sight reading, reading comprehension, math) which can be compared with other neuropsychological skills and performances on related tests. Assistance in the diagnosis of learning disabilities can also be provided by these measures to the extent that specific deficiencies are identified.

(i) *Wide-Range Achievement Test-3rd Edition (WRAT-3)*. The WRAT-3, like its predecessors (WRAT, WRAT-R), provides for an assessment of written spelling, written mathematical, and sight-word reading skills (Wilkinson, 1993). It enjoys widespread use across the lifespan (age 5–75) and is relatively brief (i.e., approximately 20–30 minutes).

(ii) *Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R)*. This is one of the most comprehensive batteries when used in its entirety for the individual assessment of global cognitive skill, general information, academic achievement, and interests (Woodcock & Mathier, 1989). It has been used with children to adults, ranging in age from two to 90 years. The WJ-R Achievement Supplemental Battery provides a more focused assessment of academic

Table 5 Dementia Rating Scale (DRS) and level of dementia.

115–130	Mild
100–114	Moderate
90–99	Moderate to severe
< 90	Severe

achievement through the use of subtests that assess reading (both nonsense words and for meaning), math concepts, proof-reading, writing to dictation, and writing/sentence composition. Various cluster scores pertaining to different aspects of achievement and level of competence across domains can be examined in order to establish a profile of abilities.

(iii) *Peabody Individual Achievement Test-Revised (PIAT-R)*. The PIAT-R is used primarily for children, but offers norms from kindergarten through high school (up to age 19). It comprises the following subtests: general information, reading recognition, reading comprehension, mathematics, spelling, and written expression (Markwardt, 1989). Various subtests may be administered in order to screen for specific areas of difficulty, and because it is infrequently assessed by many other tests of cognition, the reading comprehension subtest is perhaps one of the most commonly used subtests of the PIAT-R as part of the adult neuropsychological evaluation.

4.11.4.3 Executive Functioning, Problem-solving, and Reasoning

This group of abilities is referred to by various names, and encompasses a variety of skills and functions. Judgment, reasoning, problem-solving, and abstraction are some of the more common skills within this domain that involve some of the most complex and integrative functions of the human brain. While such abilities are not specifically localized and rely upon the integration of various cortical and subcortical regions and pathways, the frontal lobes tend to play a relatively prominent role in many of these skills.

Formal neuropsychological evaluation represents the most efficient means of assessing these functions, although more informal administration of similarities, proverb interpretation, inhibition, and reasoning and judgment tasks also can be useful. Fluency tasks have also been shown to be useful in assessing executive or anterior cerebral functions—letter or phonemic fluency (“tell me as many words as you can think of that start with ‘R’ within a minute”) to assess left anterior hemispheric functions, and design fluency for anterior right hemispheric function (e.g., having patients draw as many unique nonreplicated designs as possible within a minute). Such tasks may be modified to further tax executive systems, for example, by alternating between different procedures (“name an animal, then a city”), although normative data on such procedures are lacking.

Judgment or reasoning tasks include questions such as “What would you do if you were

stranded at the airport with only one dollar,?” or “Why should people visit the doctor when they are not ill?,” or “The tiger was eaten by the Lion—who did the eating?” As with some other tests of higher cognitive function, proverb interpretation involves a variety of skills that cannot be readily distilled into a specific set of abilities. Administration of familiar (e.g., “A rolling stone gathers no moss”) as well as unfamiliar proverbs (e.g., “The used key is always brightest”) is recommended, however, since an individual’s ability to interpret novel material may rely upon different skills compared to their ability to retrieve the common response to a more well-rehearsed task. Information about patients’ planning abilities is also useful, whether assessed by a formal cognitive task or simply by inquiring about their plans for the future given their current situation. Malloy and Richardson (1998) outline a number of bedside and formalized procedures for the assessment of frontal/executive functioning. In evaluating results from various tests of executive abilities or frontal lobe function, it must be kept firmly in mind that such measures are not uniformly or invariably affected by damage to frontal systems (Bigler, 1988; Stuss & Benson, 1984), and that normal test results can be seen in some patients with clear evidence of frontal damage, just as impairment on the same tests can occur in patients with pathology outside the frontal lobes.

4.11.4.3.1 Tests of reasoning and problem-solving

(i) *Wisconsin Card Sorting Test (WCST)*

This is a popular measure of nonverbal reasoning and cognitive flexibility (Heaton, 1981; Heaton, Chelune, Talley, Kay, & Curtis, 1993). It requires subjects to sort a series of colored geometric designs into piles according to specific principles. Feedback is given as to whether each individual response is correct or incorrect, although subjects are not instructed as to what strategies they should be using. The WCST is sometimes touted as a measure of “frontal” functioning. Whereas the test has been shown to be relatively more sensitive to anterior brain damage in some studies (Heaton, 1981), it is important to keep in mind that the WCST is not exclusively a measure of frontal functioning, and can be sensitive to damage in other areas of the brain. Thus, impairment on the WCST in and of itself cannot necessarily be interpreted to suggest disruption of frontal brain systems. Nevertheless, patients who are disinhibited and have difficulty monitoring their responses will tend to be perseverative in their responses to this

task. Furthermore, given the appropriate supporting pattern of anterior dysfunction based on behavior and results from other measures, results and observations from the WCST may assist in determining whether or not there is evidence of frontal brain system involvement. The most sensitive score from the WCST has proven to be the number of perseverative responses (Heaton, 1981), as many brain-injured individuals will tend to repeat the same strategy, despite corrective feedback. The ability to shift response strategies and perform such trial-and-error learning may also be a good predictor of various aspects of everyday functioning.

(ii) *Category Test*

The Category Test (Reitan & Wolfson, 1993) is one of the more sensitive measures of cerebral dysfunction, probably because of its complexity. This task requires subjects to generate and implement various problem-solving strategies across its 208 items. It comprises seven subtests, most of which require different principles to arrive at correct solutions. The subtests are generally arranged in increasing levels of difficulty, although two of the subtests actually require the same solution strategy, and the final subtest relies more upon memory, insofar as subjects must recall which response was correct when they saw that item in a previous subtest. The Category Test tends to be sensitive to dysfunction regardless of primary site of damage, although executive function impairment tends to have a particularly disruptive effect on results. Several versions of the Category Test are available (see Spreen & Strauss, 1998), including the standard machine version, a booklet version, a pamphlet version, and several computerized adaptations of the original test. Various short forms and child versions are also available.

(iii) *Delis-Kaplan Executive Function Scale (DKEFS)*

This is a battery of executive function tasks based on the Boston process approach that is being published by The Psychological Corporation (Delis, Kaplan, & Kramer, in press). Among other measures, it includes the California Sorting Test (CST), a measure designed to delineate more qualitative and multidimensional aspects of problem-solving, concept generation, and mental flexibility than other sorting tasks. The test consists of a series of cards that differ in shape and color and have different words printed on them. The cards can be sorted according to any of eight different

principles. Initially, subjects are asked to sort the cards into two equal groups, without any further instruction. This procedure is repeated until no new groupings are performed. After each sort, subjects are asked to indicate what rule or principle they used. Subsequent trials include stating the sorting principle used by the examiner, and sorting in different ways according to cues. The test provides for an assessment of verbal and nonverbal concept formation and perseveration and allows for information to be derived regarding which component process may be principally affected. Preliminary studies of the CST have been promising in terms of the test's clinical utility (e.g., Beatty, 1993; Delis, Squire, Bihrl, & Massman, 1992; Greve, Farrwell, Besson, & Crouch, 1995). Also, because of the flexibility of the procedure, it is appropriate for a variety of ages, and normative data on individuals aged eight to 89 will be available.

4.11.4.4 Arousal and Orientation

Evaluation of these basic functions is a necessary prelude to the examination of higher level cognitive skills. These abilities represent some of the basic building blocks upon which higher abilities rely. For example, memory cannot usually be adequately assessed in a patient who is unable to attend to stimuli. Arousal refers to basic level of consciousness as a function of the reticular activating system. Arousal is often described as ranging from alert, drowsy, lethargic, stuporous, to comatose. Alternatively, hyperarousal can also be seen in terms of restlessness, agitation, and delirium. The delirious patient is confused and transiently disoriented. Patients may seem to come "in and out of it" from day to day or hour to hour, and mental status examination results may fluctuate accordingly. Assessing level of arousal is a clinical judgment, and often, if there are difficulties in this area, a behavioral description of what it takes to arouse a patient can be most important (particularly since a more detailed examination of cognitive abilities may not be possible).

Orientation is typically thought of in terms of the three general spheres of time, place, and person. Orientation to time is the most readily disrupted of these and can be assessed by asking patients questions about the current time, day, year, month, and date. Assessment of place is obvious (i.e. "Where are we now?"), although questioning about situational variables ("Why are you here?") is also useful along these lines. Orientation to person is the most resistant to disrupted cerebral functioning relative to time and place, and refers to a patient's ability to

provide personal identifying information as well as being able to identify others who are familiar to him or her.

4.11.4.5 Attention/Concentration

The concept of attention is complex (e.g., see Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991). As some of the more basic cognitive abilities, attention and concentration are required for adequate performance on essentially all other cognitive tasks. Thus, it should be kept in mind that if a severe attentional impairment is present, associated difficulties should be manifest on tests of other cognitive abilities as well. General or undirected attention is related to overall arousal level and one's awareness of the environment. Additional divisions of attention into immediate vs. sustained attention also has clinical merit, as some patients can perform simple tasks of immediate attention, yet may have difficulty when asked to maintain their concentration longer. Measures of sustained attention, often referred to as "concentration" tasks, include measures designed to focus an individual's attentional skills for a more extended period of time, for example, over the course of a minute or longer. Informal assessment procedures may include having the examiner say a long series of random letters or numbers and having the patient indicate whenever they hear a designated target number or letter (e.g., the "A" test of attention; Strub & Black, 1988, p. 58).

Although it reflects a "global" brain activity, attention has some focal and lateralized correlates. Subcortical systems such as the reticular activating system in the core of the brainstem and midbrain, as well as diencephalic structures (e.g., thalamus) play a central role in the brain's ability to maintain arousal and focus attention. In terms of laterality issues in the visual domain, the left hemisphere attends to right hemispace and the right hemisphere attends to left hemispace. Hemi-attentional disorders, less severe than hemispatial neglect, may require double simultaneous stimulation to elicit, although evidence of hemi-inattention can also be detected in many cases by using line bisection tasks, number or symbol cancellation tasks, or even simple drawings to command or copy. As an extreme example, the patient with hemi-neglect may ignore one-half of their visual world, denying that their arm or leg belongs to them upon questioning, or eating only the food on one side of their plate. Hemi-inattention and neglect phenomena most often occur on the left side of hemispace, and the right hemisphere has been implicated as perhaps playing a predominant role in some aspects of attention. One

simple test of neglect involves line bisection, wherein the examiner draws a series of horizontal lines of different length randomly on a page. The task for the patient is to bisect each of these with a small mark. Patients with neglect will show a tendency to skew their marks to one side (i.e., away from the neglected side) of each stimulus. A wide array of standardized clinical and experimental assessment techniques is available to measure aspects of attention and concentration, and only a few of these will be discussed.

4.11.4.5.1 Assessment of attention

(i) Digit Span (WAIS-III, WMS-III, WAIS-R, WMS-R)

The familiar Digit Span task requires subjects to repeat series of digits of increasing length. Digit span forward is a good measure of simple attention, and most healthy individuals perform within the seven plus/minus two span of apprehension range. While some consider the digit span test to be representative of a memory task, the demands on memory *per se* are minimal, and digit span forward is best considered as a measure of attention. To illustrate this principle, amnesic patients with Alzheimer's disease and Korsakoff's syndrome often demonstrate a normal digit span forward despite severe anterograde amnesia. Digit Span backwards represents a qualitatively different type of task that relies more upon working memory skills that should be considered separately from digits forward (Reynolds, 1997).

(ii) Digit Symbol (WAIS-III, WAIS-R)

This is a complex measure of attention, working memory, visual scanning, and psychomotor speed. Subjects are required to learn to associate numbers with symbols and must fill in the symbol that is paired with each of nine numbers under time constraints (90 seconds). As noted, it tends to be the most sensitive of the WAIS-R subtests to brain dysfunction, and is often used separately from the Wechsler scales.

(iii) Symbol Digit Modalities Test

This is very similar to the Wechsler Digit Symbol test, in that it requires the written substitution of numbers and symbols. In this test, however, subjects must fill in the specified number that goes with the indicated symbol (Smith, 1982). An oral version is also available wherein the patient responds verbally with the number associated with each symbol they come to. Ninety seconds are allowed in this test as well, and it has been used in a large number of

neurobehavioral studies since its inception. In some research settings to assess medication effects, subjects complete the test repeatedly prior to the beginning of the study in order to decrease subsequent practice effects.

(iv) *Digit Vigilance Test*

This represents a popular and well-normed (Heaton, Grant, & Matthews, 1991) version of the traditional number cancellation task. In this test, subjects must cross out all of a specified number they come to on pages that are filled with rows and columns of single digits (Lewis & Kupke, 1977; Lewis & Rennick, 1979). The task requires visual scanning, psychomotor speed, and sustained concentration, insofar as it typically takes several minutes to complete the two-page version of this task. Additional information regarding lateralized inattention can also be derived from this and similar tasks.

(v) *Visual Search and Attention Test (VSAT)*

This provides similar information to Digit Vigilance regarding sustained visual concentration, but includes verbal and nonverbal stimuli (Trenerry, Crosson, DeBose, & Leber, 1990). The VSAT consists of four tasks (two for practice) that require subjects to cross out target items from 10 rows of 40 stimuli (letters or numbers) on each version. Scoring for left- vs. right-sided errors in addition to an overall attention index is included.

(vi) *Bells test*

This is a symbol cancellation task similar to Digit Vigilance, although the target in this case is a small picture of a bell that is repeated within a quasirandom array of 315 small visual figures on a page (Gauthier, Dehaut, & Joannette, 1989). One of the advantages of this task can be that patients who utilize more systematized visual search strategies (e.g., scanning left to right and from row to row) on more symmetrically organized tasks such as Digit Vigilance may show impairment more readily on the more randomly arranged bells.

(vii) *Continuous Performance Test (CPT)*

Several versions of this computerized test of sustained concentration have been developed, and variations on the basic task have been used in research and clinical settings since the 1950s. In one common format, the CPT presents several series of single visual stimuli (usually letters) projected at brief intervals. Subjects must respond when a letter was preceded by a

designated target letter. In Connor's CPT (Connors, 1995), subjects are asked to press the appropriate key for any letter presented, except for the letter S. Multiple trials of different interstimulus intervals are presented, and scoring includes reaction times, number correct, and errors of omission and commission.

4.11.4.6 Language

Language is a central part of the human experience, and a careful evaluation of the core components of language function are an important part of the neurobehavioral examination. Because of its complexity and multifaceted nature, an assessment of "language" requires multiple tasks. Note that some components of language (e.g., basic comprehension and expression) are amenable to observation during even casual interactions with patients, as well as during most of the traditional mental status examination. Recall that the left hemisphere is dominant for language functions in most people (particularly in right-handers, wherein over 95% are left-hemisphere dominant for language), although a minority of individuals (particularly left-handers) have bilateral speech representation or may even have their primary language center in the right hemisphere.

In screening for language disturbance, several primary areas should be assessed: verbal expression, naming, repetition, comprehension (oral and written), reading, and writing. Depending upon the function disrupted, performances on tasks involving these skills may provide useful information in terms of localization of dysfunction. Anterior dominant hemisphere systems are involved in expressive language (e.g., Broca's area), while more posterior regions are involved in receptive language (e.g., Wernicke's area). Aphasic disorders can occur as a result of a disruption of either of these areas, interconnecting cortical and subcortical pathways, or in cortical association areas nearby. The major aphasic syndromes are listed in Table 6.

Several comprehensive batteries for assessing aphasia have been developed, including the Boston Diagnostic Aphasia Examination (BDAE; Goodglass & Kaplan, 1987), the Multilingual Aphasia Examination (MAE; Benton & Hamsher, 1989; Benton, Hamsher, Rey, & Sivan, 1994), and the Western Aphasia Battery (Kertesz, 1982). Routine neuropsychological evaluations vary widely in terms of the depth with which language is examined. At a basic level, to begin to assess language function, the following areas and procedures should be considered.

Table 6 Common aphasic syndromes and abilities affected.

<i>Aphasia Type</i>	<i>Fluency</i>	<i>Naming</i>	<i>Repetition</i>	<i>Comp.</i>	<i>Rdg. comp.</i>
Broca's	Poor	Poor	Poor	Good	Good
Wernicke's	Good	Poor	Poor	Poor	Poor
Conduction	Good	Poor	Poor	Good	Good
Global	Poor	Poor	Poor	Poor	Poor
Transcortical motor	Poor	Poor	Good	Good	Good
Transcortical sensory	Fluent	Poor	Good	Poor	Poor
Mixed transcortical	Poor	Poor	Good	Poor	Poor
Anomic	Good	Poor	Good	Good	Poor

(i) Expressive language/fluency. Spontaneous speech can be examined during clinical interview and history-taking. This includes an assessment of rate of speech, prosody, grammatical complexity, and use of vocabulary.

(ii) Confrontation naming refers to the ability to name objects or representations of objects. This can be assessed in a rudimentary way by having the patient name things in the room or things you are wearing or holding. A much more sensitive means of assessing naming ability is to show the patient pictures of objects and request that they say what it is. Clinical experience suggests that the further the representation is from real life (e.g., a line drawing vs. a photo or real object), the more sensitive the task will be to dysnomia. An important note is that some degree of dysnomia is seen in most types of aphasia.

(iii) Repetition is examined by having the patient repeat single words (e.g., "paper," "ball," "comb") and phrases of increasing length and complexity (e.g., "The boys played baseball," "The vat leaks," "He is the one who did it"). Careful notation of the specific errors made (e.g., omissions, paraphasias) should be made, as these may merit further exploration using more standardized tasks.

(iv) Comprehension is also assessed in interview and during the performance of tasks upon request by the examiner (i.e., does the patient seem to understand what is being asked, or do tasks require additional explanation?). Examples of comprehension tasks include: "Point to the ceiling, then to the floor" "Do two pounds of flour weigh more than one?" "Place the quarter to the right of the penny and pick up the dime."

(v) Reading is examined readily by having the patient read single words, something they have written, and/or read sentences written by the examiner. Reading comprehension can also be tested by having the patient read a passage from a newspaper or magazine and then asking them questions about what they read (a number of standard brief tasks are available).

(vi) Writing is tested by having the patient write their name (a highly over learned task),

specific words or a sentence to dictation, or perhaps a sentence or paragraph on a topic of interest (which also allows for an assessment of sentence structure, punctuation, and grammar use, etc).

The assessment of general expressive and receptive language skills encompasses a large array of tests and procedures, and is well beyond the scope of this section. Nevertheless, various language skills should be sampled in the comprehensive neuropsychological evaluation, at least at a cursory level. This may entail the administration of specific tests (e.g., the Token Test to evaluate comprehension), or may reflect more of a basic examination of an individual's ability to repeat, speak, write, and comprehend. Various types of linguistic errors can be seen during speech and these various tasks and should be noted when they occur. Paraphasias can be semantic or verbal (i.e., the substitution of an incorrect word), phonemic or literal (i.e., the incorrect substitution of phonemes within a word or a similar sounding word), or neologistic (i.e., made-up, nonsense words).

4.11.4.6.1 Global language assessment

(i) The Aphasia Screening Test

This test is a popular and extremely brief measure that employs single- or several-item sampling of a variety of language skills (e.g., reading, writing, naming, comprehension) (Reitan & Wolfson, 1986). For tests such as this, it must be kept in mind that a single error on a particular item, while potentially significant, may not reflect an underlying deficit. Instead, this may reflect the brief nature of the task and its sensitivity to response variability during a neuropsychological examination. Hence, individual errors on such tasks should be followed up with additional tests, lest an overinterpretation of results occur.

(ii) The Wechsler Vocabulary Subtest

This is a subtest of the Wechsler scales (WAIS-III, WAIS-R) and provides a good

assessment of vocabulary and basic word knowledge. As previously noted, the Vocabulary subtest tends to be relatively insensitive to acquired cerebral damage in most cases, and may provide an index from the Wechsler scales regarding an individual's likely level of pre-morbid functioning. Unlike the NART, however, the Vocabulary subtest, with its reliance upon subject-generated word definitions, tends to be somewhat more susceptible to the effects of brain injury, and may also be impacted by conditions that affect initiation of responding. The WAIS-R-NI version includes provisions for multiple-choice responding for those patients who may have difficulty with verbal expression. The correlation between WAIS-R Vocabulary and FSIQ is 0.81 (0.85 with VIQ), and the WAIS-III Vocabulary correlation with FSIQ is 0.80 (0.83 with VIQ), also makes the Vocabulary subtest appealing when a brief estimate of current verbal intellectual functioning is desired.

(iii) *Boston Naming Test (BNT)*

This is a popular quantitative measure of confrontation naming ability (Kaplan, Goodglass, & Weintraub, 1983). Because of the prominence of dysnomia in the various aphasic syndromes, the assessment of word-finding has become a component of many neuropsychological test batteries. The standard BNT consists of 60 line drawings of objects of decreasing familiarity (e.g., ranging from "comb" to "abacus"). Several short forms of the BNT are also available (Franzen, Haut, Ranking, & Keefover, 1995; Lansing, Ivnik, Cullum, & Randolph, in press) for use when a more brief assessment of naming is desired. As noted, clinical experience indicates that the ability to name drawings of items is a much more sensitive means of assessing dysnomia than having patients name articles that are physically present in the room.

(iv) *Verbal fluency (letter and category)*

Clinical tests of verbal fluency come in two primary varieties: letter or phonemic fluency and category or semantic fluency (Benton & Hamsher, 1967; Borkowski, Benton, & Spreen, 1967). For letter fluency, subjects are asked to generate as many words as possible in 60 seconds that begin with a specified letter. Norms and alternate forms are available (Kozora & Cullum, 1994; Spreen & Strauss, 1998), although the most common sets of letters in use today include F, A, S, and C, F, L, with P, R, W serving as an alternate form. Category fluency is most commonly assessed by having

subjects generate as many animals as they can think of in 60 seconds, with other variations of this task including fruits, vegetables, first names of people, supermarket items, cities, and US states.

Performance on letter (phonemic) and category (semantic) fluency tasks has been analyzed in number of populations, and it has been shown that the abilities involved in each task are dissociable. For example, in patients with Alzheimer's disease, it has been commonly reported that patients show a greater difficulty on category fluency, and often do somewhat better on letter fluency. This is thought to be related to the breakdown in semantic knowledge as seen in patients with Alzheimer's disease, unlike patients with subcortical dementias (e.g., see Chan et al., 1995). Along these lines, Moscovitch (1994) provides evidence to suggest that letter fluency is more of a frontally-mediated task, while category fluency relies more upon temporal lobe structures.

4.11.4.7 Visuospatial Skills

This cognitive domain refers to the ability to process visual or nonverbal information. In the most common visuospatial task, patients are asked to draw a specific shape (e.g., intersecting pentagons). More difficult tasks include drawings of three-dimensional figures (cube, house). Drawings to command can be compared with patients' copies of drawings to see if the presence of a model facilitates performance. Three-dimensional constructional tasks also exist, such as replicating designs using blocks. These and related tasks require various levels of visuo-perceptual and visual-integrative skills, in addition to a motor output (e.g., graphomotor or manipulospatial).

Various standard tasks and stimuli are available for clinical use. Prominent visuospatial errors can be readily detected on such tasks, most often in association with right hemisphere damage (primarily parietal dysfunction). Lateralized errors in drawings, embellishment or paucity of creations, perseveration, rotations, intrusions, and loss of *gestalt* or detail can all be observed and may provide useful diagnostic and descriptive information. For example, patients with right hemisphere damage will tend to have more difficulty maintaining the overall *gestalt* of designs, whereas left hemisphere dysfunction more often results in problems recreating details of figures. Because many of these tasks require visual perception and a motor response, the clinician must use caution in interpreting visuoconstructional impairments when primary visual, perceptual, or motor disturbances are present.

4.11.4.7.1 Visuospatial tasks

A variety of drawings are commonly used in neuropsychological settings to assess graphomotor visuoconstructional ability. Some drawings have gained widespread popularity such as clocks, drawings of crosses, and three-dimensional cubes. A dissociation can be seen in some patients between their ability to spontaneously draw a specified design or item to command and their ability to reproduce it from a model presented to them. Obviously, the production of any drawing involves a variety of skills, and an interpretation of impaired visuoconstructional ability should only be done once a thorough assessment of the various component processes has been undertaken. For example, patients with basic motor or movement disorders will often produce impaired drawings, which should not be interpreted as a visuoconstructional deficit if their level of motor impairment can just as easily explain their poor productions.

The term "constructional dyspraxia" is often used to describe impaired visuoconstructional abilities. Unfortunately, the term is often misused and overused in clinical practice. Technically, the term refers to an impairment in the ability to construct visuospatial figures or drawings, whether two- or three-dimensional. Thus, a patient who demonstrates an impaired copy of a key from the aphasia screening test should not be considered to have constructional dyspraxia if they show normal performances on other visuoconstructional tasks such as WAIS-R Block Design, for example. Furthermore, it is important to obtain more than one drawing sample from a patient, as the degree of complexity of one stimulus may fail to elicit an impairment, whereas another one of greater complexity or less familiarity may reveal evidence of difficulty in this domain. The clinician's familiarity with developmental and general norms on a given drawing procedure are critical. This is particularly important on drawing tasks that do not lend themselves to highly detailed scoring procedures. Even when such scoring procedures are available, clinical interpretation based on a careful review of the component processes and the actual drawings produced can be critical. For example, the Rey-Osterrieth complex figure has seen the development of a variety of scoring systems over the years (see Lezak, 1995; Spreen & Strauss, 1998). Although more recent attempts such as the Boston qualitative scoring system (Stern et al., 1994) have attempted to capture more of the essential details and qualitative aspects of the design, quantitative scoring alone may prove inadequate in the absence of careful inspection by a seasoned clinician.

(i) Clock drawing test

This is a staple for many clinicians across disciplines involved in mental status examinations and neurobehavioral assessment, and actually can provide a wealth of information regarding a variety of cognitive skills. First, the reproduction of the face of a clock from memory requires some familiarity with such a stimulus. In some cultures, this represents a highly overlearned icon, and thus, should be reproduced even by children without significant difficulty, providing they are able to tell time.

Typical instructions (Goodglass & Kaplan, 1987) for the clock drawing test are to "Draw the face of a clock showing the numbers and the hands set to ten after eleven." The additional instruction to include all of the numbers on the clock can be helpful in some cases, particularly since some clocks and watches do not include a complete set of alphanumeric symbols. The instruction to set the hands at a specified time is also critical, and even when gross visuoconstructional abilities and the overall *gestalt* of the clock may be correct, time-setting errors may reflect problems with executive functions and/or memory. It is also important that the instructions include setting the hands to a time that is not too readily obvious or overlearned, so as to reduce the likelihood of an individual getting the time correct by chance. Various scoring procedures for clock drawings have been established (e.g., see Freedman, Leach, & Kaplan, 1994; Kozora & Cullum, 1994; Royall, Cordes, & Polk, 1998), although careful visual inspection and qualitative analysis, in addition to the application of a particular preferred scoring system, is important from a clinical interpretive perspective.

Along these lines, some of the more common errors that can be seen on the clock drawing test include confusion, gross distortion, and constructional dyspraxia (Figure 9(a)), aspects of unilateral neglect (Figure 9(b)), stimulus-boundness and perseveration (Figure 9(c)), and errors in hand-setting (Figure 9(d)). Figure 9(d) was drawn by a patient with mild Alzheimer's disease and demonstrates a common conceptual stimulus-bound response by their setting of the hands at the 10 and 11 in order to depict "ten after eleven." Comparison of clock drawings done to command and under copy conditions have also demonstrated utility in some differential diagnostic situations (e.g., Libon, Malamut, Swenson, Sands, & Cloud, 1996).

(ii) Aphasia Screening Test

Although designed primarily as a brief screening measure of gross language disturbance, the aphasia screening test contains several drawings,

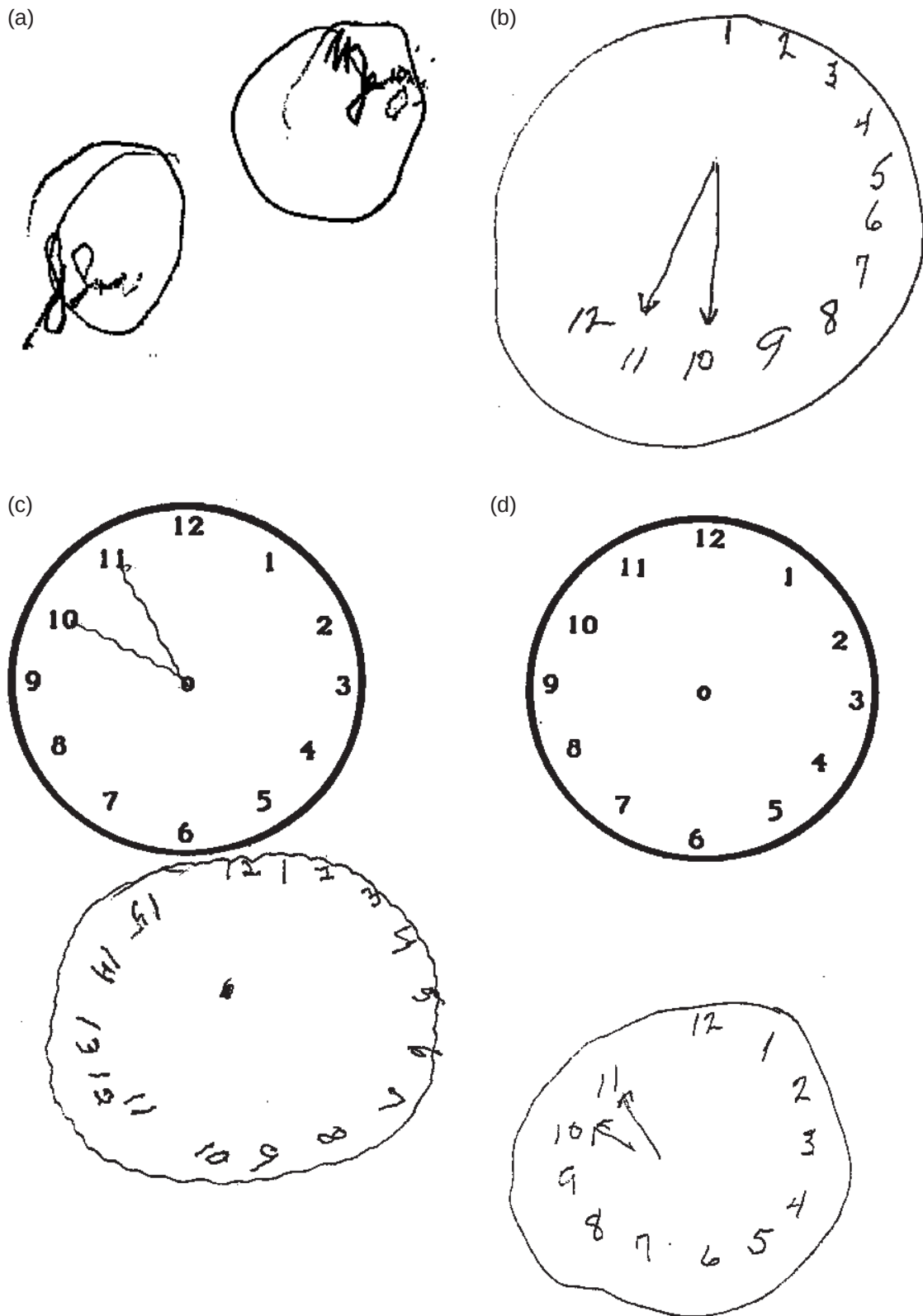


Figure 9 Clock drawings from patients showing various error types to command (9a, 9b) and copy (9c, 9d).

including a square, triangle, Greek cross, and a skeleton key (Reitan, 1984). Traditional qualitative scoring procedures for these drawings exist, although as with other similar tasks, visual inspection by a trained clinician is extremely important, and one needs to be familiar with the wide range of variability seen in the normal population as well. The copy of the Greek cross or similar crosses is a very common graphomotor construction task, and despite its relative simplicity, nevertheless does require a variety of skills, including visuoperception, planning, graphomotor control, and sequencing.

(iii) *Block Design*

The Block Design subtest from the WAIS-III or WAIS-R provides one of the most common means of assessing visuoconstructional ability. This familiar task requires the replication of red and white designs using three-dimensional colored blocks. As with other visuoconstructional tasks, careful observation of the patient's performance during the procedure can yield very important information, as can the qualitative nature of the designs that are produced. This serves as some of the rationale for including spaces for noting the designs produced by patients in the WAIS-III summary form, as the same score can be achieved for a variety of

reasons. To illustrate, Figure 10 depicts the designs from two brain injured patients.

Technically speaking, both productions would receive a WAIS-III score of zero, yet they obviously reflect very different underlying aspects of dysfunction. As has been discussed by Kaplan and co-workers (Kaplan, 1990; Milberg, Hebben, & Kaplan, 1996), many patients with right hemisphere dysfunction tend to break the overall square configuration of the design, as they have difficulties in processing the *gestalt* of figures. In contrast, patients with damage to the left hemisphere tend to more frequently make internal rotation errors. In Figure 10, the design on the left was, in fact, produced by a patient with a large subcortical infarct in the left-hemisphere, and the design on the right was produced by a patient with a large middle cerebral artery stroke.

(iv) *Rey–Osterrieth complex figure*

Although the Rey–Osterrieth figure is perhaps most well known as a measure of nonverbal memory, the copy of the figure provides an excellent sampling of visuoconstructional skills (Corwin & Bylsma, 1993; Osterrieth, 1944; Rey, 1941). In addition to providing data pertaining to an individual's organizational approach to the task (assuming the sequential replication of

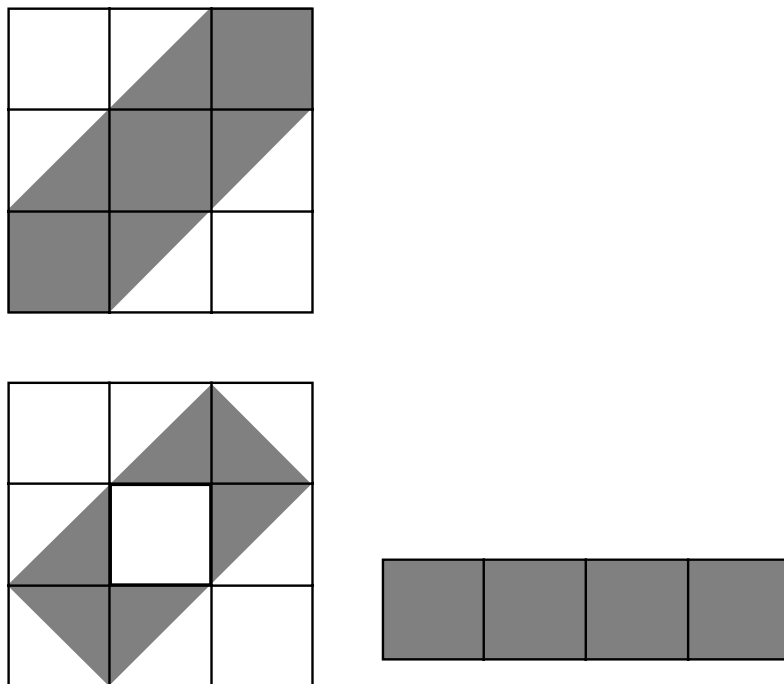


Figure 10 Block Design reproductions (top = model) of a patient with left hemisphere damage (left) and right hemisphere damage (right) showing characteristic internal detail and configural errors, respectively. Note that technically, both reproductions would receive a score of zero, although they reflect very different types of errors.

the figure is somehow carefully tracked and recorded), visuoperceptual as well as visuospatial and graphomotor constructional skills are necessary components to this task. Information pertaining to lateralized dysfunction can also be obtained in some cases (Rapport, Dutra, Webster, Charter, & Morrill, 1995). This requires not only careful attention to the *gestalt* vs. internal details of the figure, but also to left–right and quadrant by quadrant analysis, in addition to an assessment of item and figural rotations and perseverations. Gross omission or distortion of part of the figure might also implicate lateralized dysfunction. For example, Figure 11(a) depicts the Rey–Osterrieth from a patient following the rupture of a right middle cerebral artery aneurysm. The Rey–Osterrieth stimulus figure is presented in Figure 11(b).

4.11.4.8 Memory

Memory is one of the most common clinical complaints of patients referred for neuropsychological evaluation. This represents a complex set of abilities that are all too often subsumed under a generic and unitary rubric of “memory.” While multiple memory systems and skills have been identified and discussed over the years (Schacter & Tulving, 1994), some of the more useful concepts from a clinical standpoint include immediate, recent, and remote memory. Because of the multifactorial nature of memory and the use of various labels and terms to describe this set of abilities, some basic definitions of memory concepts are in order before proceeding to a discussion of common clinical memory assessment tools.

Immediate or working memory refers to a very brief memory store where information is retained for a matter of seconds before it decays, in the absence of rehearsal. A prime example is going to the phone book to look up a new number, walking across the room to the phone, and dialing the number. Unless the number is of particular relevance or is rehearsed, it will probably be forgotten immediately or shortly after the call. Evidence suggests a particularly prominent role of the prefrontal cortex in working memory.

Recent memory is the ability to encode new information and remember things learned in the recent past. Temporal lobe structures (particularly the hippocampal formation) are particularly important in this stage of information acquisition. Examples include remembering a few words or an address presented several minutes ago, or remembering what you had for breakfast. This type of memory is what is routinely assessed clinically and is what allows us to learn new information each day.

Remote memory pertains to the ability to recall information from the distant past, for example, from many years ago. Remote memory involves the recollection of previously stored data (over the course of months or years) that has been rehearsed or retrieved multiple times and in various contexts (e.g., where you went to high school, aspects of your job, stories you have learned, etc.).

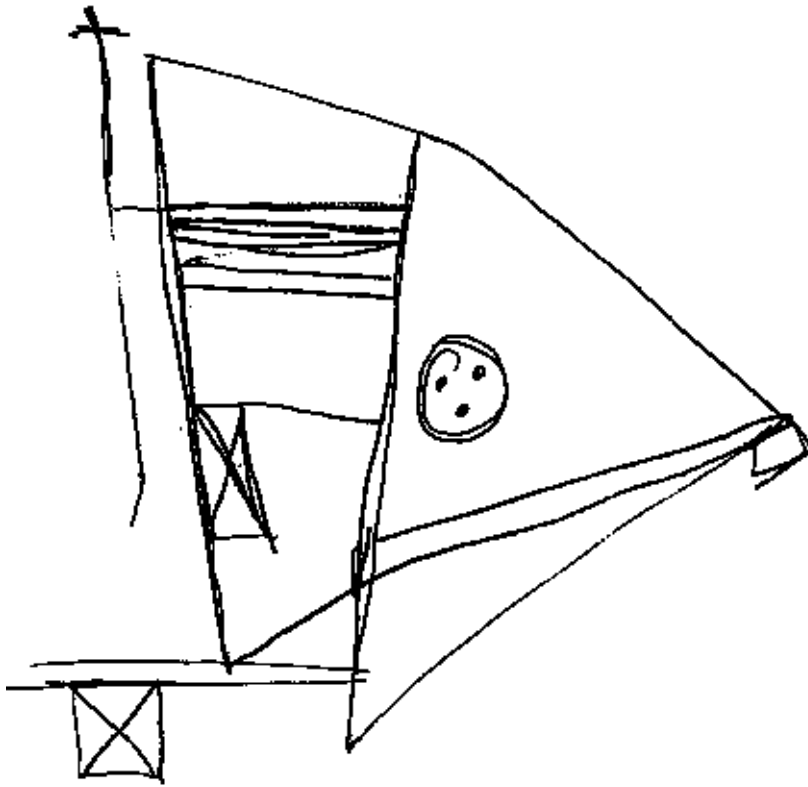
In most cases of acquired cerebral dysfunction, recent memory is the most vulnerable to damage (i.e., anterograde amnesia), while remote memory tends to remain relatively more intact. However, retrograde amnesia does occur, and often extends back into time (sometimes decades, as in Korsakoff’s syndrome) in a temporal gradient fashion, with a greater disruption of more recent memories. The clinical assessment of memory involves multiple steps and procedures. At a basic level, obtaining a history from a patient involves their remote memory abilities. Questions pertaining to familiar information historical events can also be useful (e.g., as in the WAIS-III Information subtest).

Recent memory is best informally assessed by presenting the patient with a few words, name and address, or a paragraph to recall. Here it is essential to stress that delayed recall of such information must be examined. For example, amnesic patients can often repeat back even a brief story immediately after it is presented, and it is only after additional time has elapsed (e.g., 10–15 minutes) that their full amnesic deficit may be manifest. The use of multiple memory procedures is also recommended, since significant variability even among healthy normal subjects can be seen on singular brief tests of memory. Furthermore, the use of standard or favorite memory stimuli is also encouraged (i.e., rather than inventing new stimuli for every patient), since the specific content of items can have a profound effect on their ability to be recalled (e.g. “rose, ball, key” or “apple, table, penny” are easier for most people to remember than “brown, honesty, tulip,” and these sets are no doubt easier to recall than “abstemious, benign, pusillanimous”). Regardless of what brief memory tasks are employed at the bedside, it must be kept in mind that more lengthy psychometric procedures may be required to elicit deficits in many cases.

4.11.4.8.1 Clinical assessment of learning and memory

Memory can fail for a wide variety of reasons. Problems with attention/concentration which result in an inefficiency in the ability to attend to and learn new information may be perceived as

(a)



(b)

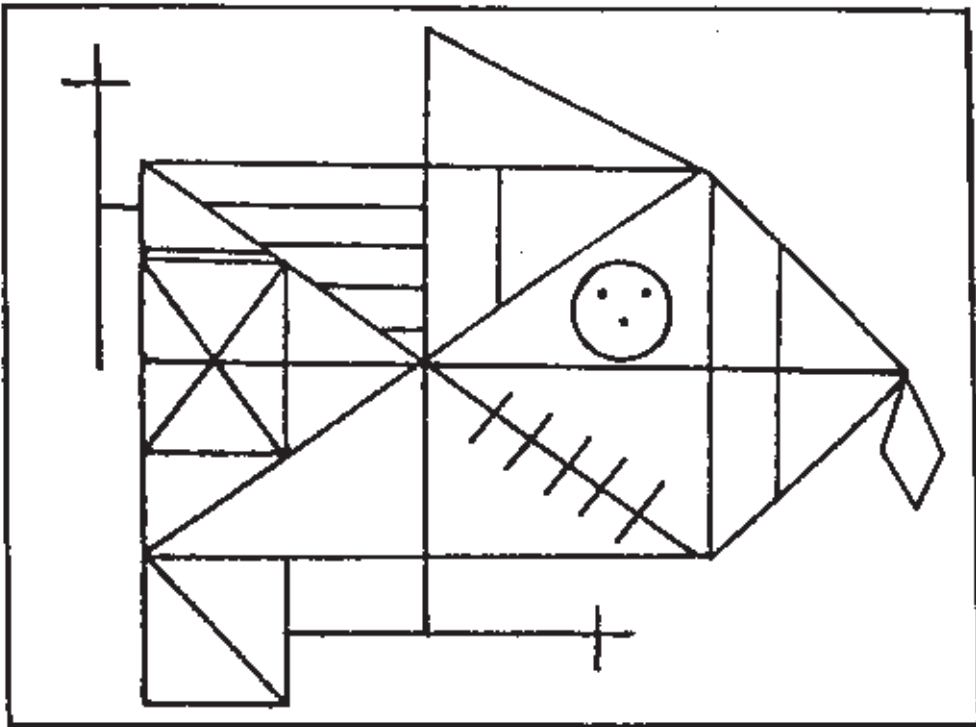


Figure 11 Rey–Osterrieth copy by a patient with right hemisphere damage (11a). Note the distortion of the left side of the figure in particular. The original Rey–Osterrieth figure is (11b).

a problem with memory, for example. If information is not efficiently encoded, it obviously cannot be stored for later recall. Subjectively, this can be perceived as a problem with remembering. Other component abilities that can interfere with efficient memory functioning include language disturbance and problems in visuospatial processing, depending upon the nature of the material that is to be remembered.

Some of the different learning and memory types that have been proposed include episodic, semantic, procedural, implicit, explicit, and working memory. A discussion of these is beyond the purpose of this chapter, although detailed elaborations can be found in sources such as Schacter and Tulving (1994) and Squire (1987). At a basic level, most of the clinical assessment approaches to memory involve an examination of episodic memory, insofar as patients are required to recall material that was presented to them earlier. Typically this material is presented in an explicit memory paradigm wherein patients are instructed to learn and remember information, although in some procedures, implicit recall is required. One example of the latter would be the typical administration of the Rey–Osterrieth complex figure, wherein subjects are asked to copy the complex design, and once that is done, they are asked to recall it from memory, without ever having been instructed to “remember” the figure.

(i) Verbal and nonverbal memory

Verbal learning and memory tasks (e.g., word lists, verbal paired associates, and story recall) have been shown to be strongly associated with left temporal lobe functioning (Chelune, 1995; Hermann, Wyler, Richey, & Rea, 1987). More debate exists about so-called “nonverbal” measures of memory and their respective sensitivity to right temporal lobe dysfunction (Barr et al., 1997). Some of this discussion centers around the nature of nonverbal stimuli that are employed in many popular memory tests. For example, many of the figures that are used are verbally encodable or lend themselves quite readily to verbal labels, thereby perhaps involving left temporal lobe structures to a significant degree. Other debate exists with regard to the specificity of functioning of the right temporal lobe, as some argue whether there is a “nonverbal” memory system at all. This is perhaps nowhere better illustrated than in cases of right temporal lobectomy in the treatment of intractable epilepsy, wherein scores on some nonverbal memory tests may show no significant change following surgery (Naugle, Chelune, Cheek, Luders, & Awad, 1993).

(ii) Forgetting rates

An important concept in memory assessment includes the examination of delayed recall for newly learned information. The simple assessment of the ability to immediately recall newly presented information, for example, may not be significantly impaired in individuals with even gross memory disorders. This can be illustrated by many patients with Alzheimer’s disease, as their ability to recall a brief paragraph that was just presented may be relatively intact, even though they show a rapid rate of forgetting over time. The seminal work by Butters and colleagues (e.g., Butters, 1985) clearly demonstrated the importance of including delayed recall techniques in the assessment of memory disorders. Along these lines, the calculation of forgetting rates or percent loss over time has proven most useful in clinical memory assessment.

(iii) Patterns of memory dysfunction

In addition to examining immediate vs. delayed recall and looking at forgetting rates, it is important to carefully analyze patterns of memory performance across individual patients. Different disorders are now known to demonstrate characteristic memory profiles in many cases, and the degree to which an individual’s profile matches what is known in the literature may be of great assistance in differential diagnostic situations. As one example, patients with Alzheimer’s disease typically show impoverished learning, rapid forgetting, faulty encoding and storage of information, and a tendency to rely upon more passive, less efficient learning strategies (Bondi, Salmon, & Butters, 1994). Their recall of stimuli that has been recently presented to them is often marked with intrusion errors (i.e., inserting features of related or unrelated stimuli during their recall of recently presented material). Like the amnesic Korsakoff’s disease patient, individuals with Alzheimer’s disease may tend to confabulate during recall in an apparent attempt to “fill in” for the material which they are no longer able to effectively organize or recall.

Another example of a memory pattern that has been shown useful in clinical assessment is the distinction between recall and recognition. Free recall involves the recollection of previously presented material. Cued recall additionally involves the presentation of a cue to help “jog” an individual’s memory. For example, if a recently presented story was about a woman who was robbed, a recall cue might be, “it was about a woman.” The next step in assisted memory is the recognition paradigm,

wherein subjects are presented with several possible stimulus items that were presented to them earlier, and the task is simply to identify which item was previously presented. It has been argued that such recognition paradigms represent “easier” tasks, insofar as the load on memory processing is diminished. An everyday analogy would be driving to a new location, being unable to remember the name of the street, but then quickly recognizing the name of the street once it is seen. Some patients with various forms of amnesia demonstrate the ability to learn, even in the absence of awareness (e.g., Schacter & Tulving, 1994), and such mechanisms may be involved in recognition memory performance.

4.11.4.8.2 Clinical memory tests

An enormous array of published measures exist for the clinical examination of memory skills, and an even larger number of tasks have been developed in the experimental and cognitive psychology literatures. A few of the tests that are commonly used in clinical practice are discussed below, and an attempt has been made to avoid too many overlapping measures. Before presenting a brief overview of individual memory assessment tools in common clinical use, one popular omnibus memory battery will be discussed. It should be noted, however, that many of the subtests which comprise the more global memory batteries are often used individually; hence, brief mention of some of the more popular subtests will be made in the following subsections.

4.11.4.8.3 Memory batteries

(i) Wechsler memory scales

This series of memory assessment batteries has a long tradition and represents some of the most popular clinical memory tests in use in the late 1990s. Because each version of the test contains various significant revisions, they will be discussed separately, although there is some overlap of core concepts and general approaches to memory testing.

(a) *Wechsler Memory Scale (WMS)*. This was the original version (Wechsler, 1945) of the popular Wechsler Memory Scales. The WMS was designed to provide a brief assessment of overall memory function by combining several tests of attention and memory. The subtests include the following: Personal and Current Information, Mental Control (e.g., counting down from 20, reciting the alphabet, counting by threes), Digit Span, Logical Memory (paragraph recall), Visual Reproduction (immediate recall of a series of geometric figures), and

Associate Learning (learning verbal associations such as “metal–iron,” “crush–dark.” A total Memory Quotient (MQ) score could be derived to permit a more direct comparison with traditional IQ scores. Thus, a significant discrepancy between MQ and IQ scores would indicate a particular impairment of recent memory. In its original version, the WMS assessed only immediate recall of verbal and nonverbal material, in addition to orientation/mental control and verbal paired associate learning. Importantly from a memory assessment standpoint, Russell (1975) added delayed recall procedures to the Logical Memory and Visual Reproduction subtests, resulting in what became known as the Russell revision of the WMS.

(b) *Wechsler Memory Scale-Revised (WMS-R)*. Administration and scoring procedures were revised and made more specific in this version of the WMS, and the floor effects of the previous test were addressed (Wechsler, 1987). Importantly, delayed recall measures were built into several of the subtests so as to assess forgetting over time. A number of new subtests to assess various aspects of memory were added, which allows for a more comprehensive assessment of different aspects of learning and memory, but also adds substantially to the time. Individual subtests include Information and Orientation, Mental Control, Digit Span, Visual Memory Span, Logical Memory I (immediate recall) and II (delayed recall), Verbal Paired Associates I and II, Figural Memory, Visual Paired Associates I and II, and Visual Reproduction I and II. The following memory index scores can be derived: Attention/Concentration, Verbal Memory, Visual Memory, General Memory, and Delayed Memory. The WMS-R was a welcome revision to the WMS, addressing some of the concerns about the earlier scale and providing much-needed improvements in standardization and norming procedures. While it has proven to be useful in differentiating various memory-disordered populations (Tröster, Jacobs, Butters, Cullum, & Salmon, 1989), the utility of the Verbal and Nonverbal Memory indices in distinguishing lateralized cerebral dysfunction is questionable (Loring, Lee, Martin, & Meador, 1989). In addition to the norms available in the test manual (age 16–74), normative data for the elderly (up to age 94) are available (Ivnik et al., 1992).

(c) *Wechsler Memory Scale-III (WMS-III)*. This represents the most recent revision of the WMS (Wechsler, 1997). The WMS-III was normed along with the WAIS-III, thereby providing for IQ and memory scores derived from the same standardization population. This allows for more direct comparisons between IQ and memory scores which should facilitate

inferences regarding the likelihood of acquired memory impairment. The standardization sample was selected to be more representative of the US population, and included individuals up to age 89. New subtests were added, and some are optional in terms of the derivation of memory index scores. Scoring procedures were further refined, updated, and expanded, and the available computer-generated scoring greatly facilitates the comparison of WMS-III and WAIS-III scores alike. Separate scores comparing working memory and immediate memory, auditory (previously referred to as "verbal") and visual (previously "nonverbal") memory, immediate and delayed recall, single-trial vs. multi-trial learning, and percent retention composites in each domain are now provided. Recognition memory procedures were also added to several of the WMS-III subtests. The subtests of the WMS-III include: Information and Orientation, Logical Memory, Verbal Paired Associates, Word Lists, Faces, Family Pictures, Visual Reproduction, Letter-Number Sequencing, Spatial Span, Digit Span, and Mental Control. Whereas this is a lengthy list of measures, five of the subtests (Information and Orientation, Word Lists, Visual Reproduction, Mental Control, and Digit Span) are supplemental or optional and do not contribute to the primary index scores. From various combinations of the core subtests, the primary index scores can be derived: Auditory Immediate, Visual Immediate, Immediate (combined), Auditory Delayed, Visual Delayed, Auditory Recognition Delayed, General Memory, and Working Memory.

The WMS-III is obviously comprehensive in its scope, now including aspects of a variety of memory assessment stimuli and procedures. It is lengthy and overlaps with some other standard memory tests in common use, but contains some interesting new procedures based on research and clinical experience in neuropsychology. Because of its recent release, studies regarding its ultimate clinical utility in various settings and populations will be needed, in addition to the development of procedures for deriving short forms and estimating memory index scores. Perhaps even more so than with the WMS-R, it is anticipated that when administration of the entire test is not feasible, many clinicians will routinely utilize selected subtests of the WMS-III to fit their particular memory assessment needs.

4.11.4.8.4 Verbal memory tests

(i) California Verbal Learning Test (CVLT)

This is a 16-item word list-learning task that contains items from four semantic categories (Delis, Kramer, Kaplan, & Ober, 1987). The list

is presented five times to subjects, in the same order, which allows for not only an assessment of learning, but of primacy-recency recall effects. The CVLT was developed with principles of cognitive neuroscience and clinical memory disorders in mind, and provides a quantitative assessment of a wide variety of qualitative features of memory performance. It has been shown to be reliable (Paolo, Troster, & Ryan, 1997) and clinically useful in a variety of clinical populations, particularly in the differential diagnostic assessment of memory disorders (Cullum, Filley, & Kozora, 1995; Peavy et al., 1994).

After the fifth presentation of the word list, an interference list is presented to permit an examination of the effects of interference on recall. Next, free recall of the initial word list is examined, followed by cued recall that involves the presentation of semantic categories for the words (e.g., "Tell me all the items that were fruits"). Long delay free recall is examined following a 20-minute delay. This is immediately followed by another series of cued recall trials, and finally, a recognition trial wherein a longer list of items containing target items as well as distractors is presented, and subjects must simply indicate "yes" or "no" whether or not these items appeared in the originally presented list. As noted, the CVLT provides for a quantified assessment of a variety of qualitative aspects of verbal learning and memory, and scoring software is available to provide normative reference scores for a host of verbal learning and memory indices. An alternate form of the CVLT has been developed that has shown good correlations with the original version (Delis et al., 1991), and a revised CVLT is in preparation in 1998. A nine-item dementia version of the CVLT has also been derived for use in more impaired populations (Libon, Mattson, et al., 1996).

(ii) Hopkins Verbal Learning Test (HVL)

This test was developed to provide an abbreviated assessment of word list-learning and memory (Brandt, 1991). It is particularly well suited for more impaired patients who might not be able to complete measures such as the CVLT. A list of 12 items from three semantic categories is presented across three trials, followed by recognition testing. A modification of the original HVL provides for an assessment of delayed recall prior to recognition in order to assess forgetting over time. A unique aspect of the HVL is that it includes six alternate forms, thereby reducing practice effects and making it a good choice in serial assessment situations.

(iii) *Logical Memory (WMS, WMS-R, WMS-III)*

This is the classic story learning and memory test that is one of the most widely used clinical indices of memory function. It is a measure of the ability to learn and retain new structured verbal information. The original WMS included two stories that were read to patients, followed by an assessment of immediate recall. The WMS-R updated the stories, norms, and scoring procedures, and importantly, added a delayed recall trial. Each of the WMS-R stories contains 25 bits of information, and a standard 30-minute delayed recall procedure provides for a ready assessment of forgetting rates. The WMS-III version includes an initial story that is administered in standard fashion (i.e., one trial, followed by immediate and then delayed recall), and a second story that is repeated in order to assess the effects of learning on storage and retrieval of material.

Patients' responses should be written down verbatim in order to glean additional qualitative information, and to allow for cross-checking of scoring. Scoring for Logical Memory is done largely based upon identical or almost-identical recall of the story material, although several scoring procedures allow partial credits to be assigned for close approximations, and the WMS-III version also incorporates scores for thematic units or gist. Logical Memory has been shown to be highly useful in the assessment of a wide array of memory disorders and sensitive to left hippocampal damage in particular (Sass et al., 1992).

4.11.4.8.5 Nonverbal memory tests

(i) *Rey–Osterrieth complex figure*

This is a very popular measure of nonverbal memory and is described and presented in several sources (e.g., Corwin & Bylsma, 1993; Lezak, 1995; Rey, 1941; Spreen & Strauss, 1998). It consists of a complex figure made up of various subcomponents. Administration procedures vary somewhat (see Meyers & Meyers, 1995), although in one of the more common administrations the task requires patients to first copy the figure, and then immediately upon its removal (and without instructions to remember the figure), reproduce it from memory. Fifteen minutes later, delayed recall is assessed. Recognition procedures have also been developed and used in some settings. Several scoring systems for the test have been developed (e.g., see Lezak, 1995; Loring, Martin, Meador, & Lee, 1990), although qualitative interpretation of patients' reproductions often provide much more detailed in-

formation than quantitative scores alone. The Boston scoring approach (Stern et al., 1994) to the Rey–Osterrieth attempts to provide a quantitative assessment of many of the qualitative features of the figure.

(ii) *Visual Reproduction (WMS, WMS-R, WMS-III)*

This is one of the most commonly used measures of visual or nonverbal memory, and is often used in conjunction with its verbal counterpart, Logical Memory, even when the WMS are not used in their entirety. Although each version of the WMS contains somewhat different stimulus figures (only two of the original WMS figures have been included in both the WMS-R and WMS-III, and Visual Reproduction is optional in the latest rendition), the administration procedures are similar. It consists of the presentation of several cards with geometric designs on each. Each card is presented to subjects for 10 seconds, and after each is removed from view, immediate recall is assessed by having the subject draw the design from memory. Following a delay of 20–30 minutes (25–35 minutes in the WMS-III), subjects are asked to reproduce the designs again from memory.

As noted, Visual Reproduction has long been used as a measure of nonverbal memory. While some literature indicates that the test does not readily lend itself to localization of cerebral dysfunction in large groups and may be insensitive to nondominant temporal lobe resection (Barr et al., 1997; Naugle et al., 1993), it can be useful in some individual cases as there may be a tendency for patients with right temporal lobe dysfunction to do more poorly (Chelune & Bornstein, 1988). Furthermore, it can be of assistance in providing a more detailed examination of memory function beyond the assessment of learning and memory for words and stories.

(iii) *Warrington Recognition Memory Test (RMT)*

This is a test of word and face recognition that allows for a comparison between basic verbal and nonverbal abilities (Warrington, 1984). First, a series of 50 single words is presented in a booklet, followed by yes/no recognition testing. The same procedure is performed for a series of 50 black and white pictures of unfamiliar male faces. While some studies of patients with lateralized brain damage have demonstrated the expected associations between left hemisphere/word recognition and right hemisphere/face recognition deficits (Warrington, 1984), this

relationship does not always hold true (Naugle, Chelune, Schuster, Luders, & Comair, 1994), and the test is perhaps best used as a supplemental test of memory in conjunction with more primary measures. In clinical practice, many patients seem to have more difficulty with the facial recognition portion of the RMT, and the association between words and faces with left vs. right hemisphere dysfunction is variable.

4.11.4.9 Motor and Sensory Function

A variety of clinical procedures have been developed in neuropsychology and behavioral neurology to examine simple and complex motor functions. These range in complexity, degree of standardization, and level of knowledge required for appropriate interpretation/analysis. At a most basic observational level, how a patient walks and moves during an evaluation can provide useful information regarding aspects of their motor functioning. Evidence of unilateral motor weakness, tremor, or other unusual movements may reflect a neurological condition or even suggest a focal cerebral deficit. These observations can also spur the astute observer to inquire more carefully about specific symptoms and may help guide the neuropsychological evaluation in terms of test selection. To take an obvious example, the patient presenting with unilateral left-sided motor weakness (e.g., foot drag and/or decreased limb mobility) should undergo careful inquiry about their symptoms, and depending upon the situation, the clinician may wish to follow up on this observation by including appropriate sensory, motor, and cognitive tasks in his/her evaluation in order to assess abilities associated with right hemisphere function.

The primary motor and sensory areas of the cortex, as well as their multiple ascending and descending fiber tracts, are well known and charted. It must be kept in mind, however, that sensory and motor regions have significant overlap in terms of cortical surface representation, and deficits are often multimodal in nature, with more complex integrative functions having a broader distribution. In addition to the assessment of two-point discrimination and simple touch, for example, more subtle cortical deficits may be detectable by the use of double simultaneous stimulation techniques. These are typically conducted in the visual, auditory, and haptic modalities, and a variety of simple quantifiable procedures can be used. Double simultaneous stimulation in the tactile modality involves repeated trials of unilateral light touch

to the back of either hand, interspersed with bilateral touch. Subjects are asked to indicate which hand was touched. Tactile extinctions or suppressions occur when one side is not perceived, and if consistent, this may suggest contralateral parietal lobe dysfunction.

Basic motor function can be quantified using various techniques and instruments, depending upon the degree of detail that is desired about a given function. Simple finger tapping speed can be assessed crudely by having the patient tap their first finger to their thumb as quickly as possible, first with the dominant, then non-dominant hands. Grip strength can be assessed by having the patient squeeze the examiner's fingers, but this technique tends to be extremely crude and is better left to a dynamometer when accurate information is desired. Fine motor dexterity can be assessed using a variety of procedures, including having the patient touch each finger consecutively to their thumb in rapid, repetitive fashion. Stereognosis can be readily examined by having the patient name objects placed in their hand without the aid of vision.

Whatever motor or sensory assessment task used, is important to note that noncortical or noncentral nervous system (non-CNS) (i.e., subcortical, spinal, peripheral) damage can produce deficits in motor and/or sensory function. Furthermore, other factors such as inattention or impairments in concentration or memory may result in poor scores or errors on sensorimotor tasks. In the absence of neuropathology, gross errors or variable responses should alert the clinician to the possibility of poor effort. In practice, motor and sensory deficits tend to be consistent, such that repeated trials later in the examination should yield similar results. Interpretation of results from these tests should always begin with these considerations in mind.

4.11.4.9.1 Psychometric measures of motor and sensory function

(i) Finger Tapping Test

This test is a simple assessment of index finger tapping speed using a mechanical tapping key device (Reitan & Wolfson, 1993). Patients are instructed to tap using their dominant hand index finger as fast as possible, and the number of taps performed within a series of 10-second intervals is recorded and averaged. These results are compared with similar scores from the nondominant hand. Most right-handed individuals tend to show approximately a 10% dominant hand superiority. Gross aberrations from this may reflect contralateral anterior

cerebral dysfunction, although it is important to note that arthritis or peripheral injuries to the brachial plexus, arm, hand, or fingers can have profound effects on this test and must be carefully ruled out before cerebral implications are made.

(ii) *Grooved Pegboard*

This is a test of fine motor dexterity that requires subjects to pick up small (one inch) metal pegs that are rounded on one side and squared on the other (Kløve, 1965). The task is to pick them up one at a time and place them in similarly shaped holes on the test platform as quickly as possible. First, the dominant hand is tested, followed by the nondominant hand. The task requires fine motor coordination and speed.

(iii) *Hand dynamometer*

Various manufacturers provide grip strength meters to assess basic hand strength. This is a device that requires patients to simply squeeze a handle as hard as they can, and the amount of force they exert is indicated in pounds or kilograms on the instrument. Two or three trials are usually administered with the dominant and nondominant hands. Dominant hand grip strength usually tends to be approximately 10% stronger than nondominant in most right-handed individuals.

(iv) *Sensory-Perceptual Examination (SPE)*

This test is included in the HRB, but is often used apart from this battery in its entirety or in portions (Reitan & Wolfson, 1985). The basic functions assessed by the SPE include finger gnosis, graphesthesia, stereognosis, and basic sensation and extinctions to simple stimulation and double simultaneous stimulation in the visual, auditory, and tactile modalities. It essentially represents a quantified approach to some of the sensory testing often performed in the traditional neurological examination. In the absence of peripheral damage factors, arthritis, severe cerebral dysfunction, or prominent attentional difficulties (which can make the interpretation of patient responses more difficult), deficits associated with these various tasks may suggest some degree of impairment of the sensory cortex and/or associated pathways.

4.11.4.10 Assessment of Motivation

The sensitivity and clinical utility of neuropsychological measures is largely dependent upon the cooperation and effort of the patient

being evaluated. Poor effort can result in artificially low or otherwise abnormal neuropsychological test scores, and the clinician must always be alert to this potential confound during test administration and interpretation. Careful analysis of levels and patterns of neuropsychological performance in relation to the disorder in question is critical (i.e., "Do the results make sense given what has happened to the patient?"). For example, the patient with an uncomplicated mild traumatic brain injury should not demonstrate severe, widespread neuropsychological deficits, and this finding in such a case should raise questions regarding levels of motivation and effort. Furthermore, as serial testing frequently occurs within the context of forensic neuropsychological evaluations, the examination of consistency of scores within and across test sessions can be a very good source of information regarding the likelihood of invalidity or unreliability of results (Cullum, Heaton, & Grant, 1991; Reitan & Wolfson, 1997).

The issue of malingering in relation to neuropsychological assessment has been reviewed in several excellent sources (e.g., Franzen, Iverson, & McCracken, 1990; Nies & Sweet, 1994). Suspected malingering or willful distortion/exaggeration of deficits continues to pose challenges in various clinical settings. Along these lines, a number of specialized tests and procedures have been developed to help detect exaggerated or feigned deficits. Some of these involve the administration of simple tasks that are presented as being difficult (e.g., Rey's 15-item test; Lezak, 1995). Others that show promise are based upon the time to complete shorter or less difficult vs. longer tasks (e.g., the Dot Counting Test; Lezak, 1995; Rey, 1941). Another technique involves the comparison of unusual patterns of performance with results known to be associated with a particular deficit or disorder (e.g., Brandt, 1988). Perhaps the most popular approach has been to utilize forced-choice recognition testing (Guilmette, Hart, & Giuliano, 1993; Hiscock & Hiscock, 1989), which relies upon probability theory by identifying performances that fall below a chance level (e.g., the Portland Digit Recognition Test; Binder, 1993).

The detection of malingering poses a complex challenge to clinicians, particularly as a variety of psychological and situational factors (e.g., gain issues) may further complicate the diagnostic picture. While it is difficult to definitively prove malingering in a particular case, the existing research suggests that it is often possible to detect evidence of malingering, although this requires skillful and deliberate inquiry by the clinician.

As with diagnosis, the clinical history can prove invaluable in establishing the likelihood of malingering. Consistency of reporting is important, and information from various sources can be compared, including clinical interview data and reports from the patient to other professionals. The consistency or “fit” between neuropsychological test results and reports of everyday functioning can also be highly useful. For example, the patient who performs in the severely impaired range on formal memory testing should demonstrate difficulties with memory in their everyday functioning. Likewise, neuropsychological test data obtained from the patient who has gross difficulty with coordination, sequencing, memory, and reaction times who rides skillfully away on his/her motorcycle following the evaluation should be carefully scrutinized.

As indicated by Nies and Sweet in their review of the literature, “Note that, at present, multiple measures and methodologies are needed, in that no single measure or methodology has proven sufficient to date” (1994, p. 544). They recommend some of the following strategies that may assist in the clinical detection of malingering: (i) use specific tests of malingering as outlined above, (ii) evaluate patterns on clinical neuropsychological tests where below-chance performances can be assessed, (iii) examine unusual or aberrant test results, (iv) examine consistency of test scores within and across measures and test sessions, (v) obtain independent information regarding a patient’s reports of their everyday functioning, and (vi) obtain a detailed clinical history and carefully examine the information provided, both in terms of accuracy and consistency.

4.11.4.11 Personality and Emotional Functioning

Changes in the functional integrity of brain systems can result in alterations in emotional reactivity (e.g., lability) and personality. In some cases, subtle changes in behavior and personality may represent the primary or earliest harbingers of brain dysfunction. The evaluation of emotional or personality change associated with neurological disease or damage poses a complex set of challenges. Damage to the frontal and prefrontal cortex can be particularly associated with personality change. This may present variably as decreased responsiveness (e.g., apathy or amotivation) or increased responsiveness (e.g., hypomania, increased or indiscriminant drives), depending upon the primary site of damage. In cases of lateralized brain dysfunction, left hemisphere lesions (particularly more anterior) have been shown

to be associated with depressive symptomatology, while lesions of the right hemisphere tend to be more frequently associated with varying degrees of unawareness or denial (anosognosia) of deficit (e.g., see Lacritz & Cullum, 1998).

Right hemisphere damage can also result in aprosodia, which refers to the impaired ability to communicate meaning using intonations and inflections during speech (emotional language). Some data suggest that the aprosodic syndromes secondary to right hemisphere dysfunction parallel the aphasia (Ross, 1985), with anterior right hemisphere lesions being more associated with expressive dysprosody, and more posterior lesions associated with receptive dysprosody. Expressive prosody testing can be readily accomplished by asking the patient pretend to be an actor and say a neutral sentence such as “I am going to the movie” using different affective intonations (happy, sad, angry, indifferent). Receptive prosody can be assessed by the examiner (while outside of the patient’s view) saying a sentence in different tones of voice and having the patient indicate if the examiner sounded happy, sad, angry, indifferent, etc. The patient with expressive dysprosody will generally sound monotone and be unable to alter the affective tone of their voice across test conditions aside from perhaps becoming louder or softer at times. The patient with receptive dysprosody will be able to say the sentence with variable affective intonations, but will have difficulty guessing the affective valence of the examiner’s sentences.

The issue of changes in emotional functioning following brain damage represents a complex biopsychosocial interaction that merits careful investigation, since multiple factors may be involved. Obviously, an insult to the brain can have a direct neurobiological effect upon the way in which information and affect is processed. However, individual psychological reactions to changes in cognitive and/or physical functioning are also commonly seen. To help establish whether neurobiologically-induced changes have occurred, the clinician must be careful to obtain information regarding patients’ premorbid functioning from family members or others who know the patient, as well as patients themselves. Defensiveness or unawareness on the part of the patient and/or family may exist, and it must be kept in mind that the clinician may be hearing a biased picture of the situation. The issue of secondary gain and malingering as discussed above must also be considered, particularly in litigation situations. Finally, the potential effects of pre-existing psychopathology should be examined, and psychiatric and/or more detailed psychological evaluation may be useful along these lines. Direct questioning about symptoms and

behavioral changes obviously may be productive, although in some instances, the line of inquiry must be more subtle, requiring indirect probing of evidence for personality/emotional changes. Rephrasing questions and asking about related symptoms in a variety of contexts may also be a useful means of not only verifying information, but in some cases, uncovering important data.

Formal psychological assessment in patients with neurological disorders requires some adjustments to standard interpretations, insofar as the endorsement of certain symptoms may simply relate to the neurological condition rather than implying psychopathology. For example, patients with neurologic disease often demonstrate clinical elevations on the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) that should not be necessarily attributed to psychological disturbance. Scale 8 in particular contains a number of items dealing with unusual symptoms, and scales 1 and 3 relate to aspects of physical functioning which may be endorsed by neurological patients due to their physical status, in the absence of psychopathology. Table 7 presents some of the more commonly used measures of psychological functioning within the context of the neuropsychological evaluation.

4.11.4.12 Test Selection Issues

As noted, specialized core batteries of tests are often assembled by neuropsychologists for specific populations. Test selection is based on training and experience in combination with information obtained from the relevant literature. Table 8 presents an example of a test battery designed to assess dementia (both in terms of early detection as well as assistance with differential diagnosis and staging of disease). Note that individual measures can readily be added or dropped (i.e., a “stepdown battery”) from the “core” set of tests, depending upon specific patient needs and clinical questions.

An example of a modification to this battery in the case of a high functioning business executive referred for evaluation secondary to suspected

incipient Alzheimer’s disease might include the following: administration of a complete WAIS-III to obtain a more precise assessment of current intellectual functioning and allow for analysis of subtest patterns; possible omission of the DRS since it may not be sufficiently sensitive to subtle deficits in such an individual in the early stages of dementia; addition of the category test to provide a more challenging evaluation of higher-order cognitive abilities than the WCST; substitution of the Ravens Advanced Matrices for the coloured version, as the latter would likely be too simplistic to be of much value in such a case; administration of a complete WMS-III to provide a more thorough examination of memory function.

Alternatively, the test battery might be dramatically altered in the case of a patient with known dementia who is being followed to track their rate of cognitive decline (or perhaps to assess for any positive effects of cognitive enhancing medications). As a more extreme example, if an individual presents with clear evidence of severe dementia during clinical interview (or if the patient had a previous MMSE score of 10), it may suffice to administer the DRS and obtain only a brief screening assessment of cognitive domains (e.g., by administering all or parts of the CERAD neuropsychological test battery). In such a case, a brief neuropsychological examination may nevertheless be considered “comprehensive” insofar as a range of abilities is assessed, and the addition of multiple complex measures may actually yield little additional neurobehavioral data, depending on the individual patient and referral questions.

4.11.4.13 Relationships Between Neuropsychometry and the Behavioral Geography of the Brain

Neuropsychological procedures, like the neurological examination, are designed to provide information about the integrity of human brain function. While these measures are sensitive to impaired cognition and may provide information regarding lateralized or focal dysfunction, it

Table 7 Common measures of psychological and emotional functioning.

Minnesota Multiphasic Personality Inventory-2 (MMPI-2)	Psychopathology/emotional status
Personality Assessment Inventory (PAI)	Psychopathology/emotional function
Millon Clinical Multiaxial Inventory (MCMI)	Personality styles, psychopathology
Beck Depression Inventory (BDI)	Depression
Inventory of Depressive Symptomatology (IDS)	Depression
Profile of Mood States (POMS)	Brief review of psychiatric symptoms
Symptom Checklist-90 (SCL-90)	Brief review of physical and psychiatric symptoms

Table 8 Sample test battery for the evaluation of dementia.

<i>Global cognitive functioning</i>
Dementia Rating Scale
WAIS-III/WAIS-R IQ or IQ estimate (e.g., based on Vocabulary and Block Design)
NART-R estimated premorbid IQ
<i>Reasoning/Cognitive flexibility</i>
Wisconsin Card Sorting Test
Similarities and proverbs testing (WAIS-III)
Trail Making Test, Part B
<i>Language</i>
WAIS-III Vocabulary
Boston Naming Test
Verbal fluency (letter and category, e.g., FAS, animals)
Token Test and/or Comprehension subtest from BDAE
<i>Visuospatial skills</i>
WAIS-III Block Design
Clock drawing
Rey–Osterrieth Complex Figure (copy)
<i>Attention/concentration</i>
WAIS-III Digit Span (forward)
Trail Making Test, Part A
<i>Learning and Memory</i>
California Verbal Learning Test
Wechsler Memory Scale-3 (WMS-III) or Wechsler Memory Scale-Revised (WMS-R) or portions thereof (e.g., Logical Memory and Visual Reproduction)
Rey–Osterrieth Complex Figure (immediate and delayed recall)
<i>Motor functioning</i>
Finger tapping test
Rapid Alternating Movements/Luria three-step

should be kept in mind that the neuroanatomical specificity of most of the tests is limited (Dodrill, 1998). Furthermore, while the frontal, temporal, parietal, and occipital cortices subserve specific functions to some degree, the amount of overlapping functional representations and inter-relationships cannot be underscored. It is worth recalling that even “primary” motor and sensory regions of the cortex have overlapping cellular representations and functions to some degree. Thus, the extent to which various cognitive functions are said to be “localized” is somewhat relative, although a number of functional neuroanatomic correlations exist with various measures. Higher cognitive functions in particular tend to defy localization, although frontal systems (particularly prefrontal cortices) tend to be heavily involved in the many abilities associated with executive functioning (e.g., planning, reasoning, judgment, insight) and personality. Limbic structures of the temporal lobes play a predominant role in learning and memory for new information, with verbal memory skills being strongly associated with left hemisphere (specifically temporohippocampal) functions in most individuals. Although less well supported, association is also often seen between nonverbal memory abilities and nondominant temporal lobe functioning. The limbic system is also intimately

associated with aspects of emotional processing, and links between affective memory and limbic functioning have been shown. While functional neuroanatomic distinctions exist (by lobe as well as in the anterior–posterior and lateral planes), the myriad interconnections that exist between these areas and other cortical and subcortical systems indicate that various brain functions may be disrupted when specific circuits or systems are damaged, even when the cortical region known to be largely responsible for a given function may appear to be intact. Thus, a careful probing of multiple aspects of brain–behavior relationships is required to provide a thorough assessment of function.

4.11.4.14 Training in Neuropsychology

Because of the complexities inherent in the neurobehavioral assessment of human brain function, specialized training is needed for the clinical neuropsychologist. Neuropsychologists must have a firm understanding of basic brain organization and function, and need to be familiar with neurological disorders and neuropathology as well as their myriad behavioral sequelae. As such, specific knowledge bases must be mastered that go beyond the traditional training of the doctoral degree in psychology.

Several documents relevant to training and practice in the specialty of clinical neuropsychology have been published over the years (e.g., Adams & Rourke, 1992; Eubanks, 1997; Reports of the INS-Division 40 Task Force on Education, Accreditation, and Credentialing, 1987), although the most comprehensive and far-reaching proviso to date was published in 1997. The Houston Conference on Specialty Training and Education in Neuropsychology was held in September 1997 and outlines recommended training and education guidelines for the future in clinical neuropsychology (Hannay et al., 1998). For this conference, neuropsychologists representing various backgrounds, employment settings, and organizations were selected from applications across the USA and Canada and invited to participate. One of the goals of the Houston Conference was to develop the guidelines for an integrated model of education and training in clinical neuropsychology, and this includes the following generic definition of a clinical neuropsychologist:

A clinical neuropsychologist is a professional psychologist trained in the science of brain-behavior relationships. The clinical neuropsychologist specializes in the application of assessment and intervention principles based on the scientific study of human behavior across the lifespan as it relates to normal and abnormal functioning of the central nervous system. (p. 161)

The recommendations for coursework, training, and skills are provided in some detail, including a model of integrated education across the doctoral, internship, and postdoctoral or residency levels. In addition to core courses in clinical psychology (e.g., psychopathology, assessment, ethics), coursework in areas pertaining specifically to brain-behavior relationships are recommended, including neuroanatomy, psychopharmacology, and various aspects of neuropsychology *per se*. It is noted in the Policy Statement of the Houston Conference that the proposed guidelines were not intended to be applied retroactively or to those currently in training at the time of the conference and proceedings, but rather, to serve as recommendations for future education and training in clinical neuropsychology. Whereas the specific functions of neuropsychologists vary depending upon setting, the following seven professional activities were put forth as being important for the clinical neuropsychologist: assessment, intervention, consultation, supervision, research and inquiry, consumer protection, and professional development. A postdoctoral residency of the equivalent of two years full time experi-

ence is included in these recommendations. Some of the exit criteria for the postdoctoral residency include demonstration of advanced skills in neuropsychology, eligibility for state or provincial licensure, and eligibility for board certification in clinical neuropsychology.

4.11.4.15 Challenges for Neuropsychological Assessment

The neuropsychological evaluation represents the most sensitive existing means by which to assess the functional integrity of the human brain. Modern structural imaging provides us with wonderfully detailed images of intracranial contents, but even knowing that an abnormality exists in a given region cannot provide much insight into how that patient is able to function and process information in day-to-day life. Functional imaging has seen major advancements in recent years and is able to indicate areas of relative dysfunction—but this, too, cannot indicate what specific cognitive abilities are impaired, or to what extent. Such techniques also offer little or no insight into how to help brain-injured patients adapt and better compensate for their altered cognitive processing, as this is the purview of clinical neuropsychology.

Neuropsychological assessment is faced with the ongoing challenge of improving diagnostic and prognostic procedures and integrating findings from clinical neuropsychology, cognitive psychology, and neuroscience into assessment tools that can provide increasingly detailed examinations of some of the intricacies of human brain function. Many of the neuropsychological tests in common use in the 1990s were developed years or decades ago, and even though some procedures are “tried and true” and continue to provide important neurodiagnostic and neurodescriptive information, the field must strive to keep up with the rapid advances being made in neuroimaging and neuroscience and be open to incorporating this new information into assessment procedures.

ACKNOWLEDGMENTS

Thanks to Kathy Saine, Ph.D., and Eric Smernoff, M.A., for their assistance with this manuscript.

4.11.5 REFERENCES

- Adams, K. M., & Rourke, B. P. (1992). *The TCN guide to professional practice in clinical neuropsychology*. Amsterdam/Lisse: Swets & Zeitlinger.
- Adams, R. L., Smigielski, J., & Jenkins, R. L. (1984). Development of a Satz-Mogel short form of the WAIS-R. *Journal of Consulting and Clinical Psychology*, 52, 908.

- Baddeley, A. (1992). Working memory. *Science*, 255, 556-559.
- Barr, W. B., Chelune, G. J., Hermann, B. P., Loring, D. W., Perrine, K., Strauss, E., Trenerry, M. R., & Westerveld, M. (1997). The use of figural reproduction tests as measures of nonverbal memory in epilepsy surgery candidates. *Journal of the International Neuropsychological Society*, 3, 435-443.
- Beatty, W. W. (1993). Age differences on the California Card Sorting Test: Implications for the assessment of problems solving by the elderly. *Bulletin of the Psychonomic Society*, 31, 511-514.
- Benton, A. L., & Hamsher, K. de-S. (1989). *Multilingual Aphasia Examination*. Iowa City, IA: AJA Associates.
- Benton, A. L., Hamsher, K. de-S., & Sivan, A. B. (1994). *Multilingual Aphasia Examination* (3rd ed.). Iowa City, IA: AJA Associates.
- Bigler, E. D. (1988). Frontal lobe damage and neuropsychological assessment. *Archives of Clinical Neuropsychology*, 3, 279-297.
- Binder, L. M. (1993). An abbreviated form of the Portland digit recognition test. *The Clinical Neuropsychologist*, 7, 104-107.
- Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist*, 3, 129-136.
- Bondi, M. W., Salmon, D. P., & Butters, N. (1994). Neuropsychological features of memory disorders in Alzheimer's disease. In R. D. Terry, R. Katzman, & K. L. Bick (Eds.), *Alzheimer's disease* (pp. 41-63). New York: Raven Press.
- Borkowski, J. G., Benton, A. L., & Spreen, O. (1967). Word fluency and brain damage. *Neuropsychologia*, 5, 135-140.
- Brandt, J. (1988). Malingered amnesia. In R. Rogers (Ed.), *Clinical assessment of malingering and deception*. New York: Guilford Press.
- Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new verbal memory test with six equivalent forms. *The Clinical Neuropsychologist*, 5, 125-142.
- Brooker, B. H., & Cyr, J. J. (1986). Tables for clinicians to use to convert WAIS-R short forms. *Journal of Clinical Psychology*, 42, 982-985.
- Butters, N. (1985). Alcoholic Korsakoff's syndrome: An update. *Seminars in Neurology*, 4, 226-244.
- Butters, N., Grant, I., Haxby, J., Judd, L. L., Martin, A., McClelland, J., Pequegnat, W., Schacter, D., & Stover, E. (1990). Assessment of Aids-related cognitive changes: Recommendations of the NIMH Workshop on neuropsychological assessment approaches. *Journal of Clinical and Experimental Neuropsychology*, 12, 963-978.
- Cermak, L. S. (1994). *Neuropsychological explorations of memory and cognition*. New York: Plenum.
- Chan, A. S., Butters, N., Salmon, D. P., Johnson, S. A., Paulsen, J. S., & Swenson, M. R. (1995). Comparison of the semantic networks in patients with dementia and amnesia. *Neuropsychology*, 9, 177-186.
- Chelune, G. J. (1995). Hippocampal adequacy versus functional reserve: Predicting memory functions following temporal lobectomy. *Archives of Clinical Neuropsychology*, 10, 413-432.
- Chelune, G. J., & Bornstein, R. A. (1988). WMS-R patterns among patients with unilateral brain lesions. *The Clinical Neuropsychologist*, 2, 121-132.
- Christensen, A.-L. (1984). The Luria method of examination of the brain-impaired patient. In P. E. Logue & J. M. Schear (Eds.), *Clinical neuropsychology: A multidisciplinary approach*. Springfield, IL: C.C. Thomas.
- Conners, C. K. (1995). *Conners' continuous performance test*. New York: Multi-Health Systems.
- Corwin, J., & Bylsma, F. W. (1993). Translations of excerpts from Andre Rey's *Psychological examination of traumatic encephalopathy* and P.A. Osterrieth's *The Complex Figure Copy Test*. *The Clinical Neuropsychologist*, 7, 3-21.
- Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *Journal of the American Medical Association*, 269, 2386-2391.
- Cullum, C. M., Filley, C. M., & Kozora, E. (1995). Episodic memory function in advanced aging and early Alzheimer's Disease. *Journal of The International Neuropsychological Society*, 1, 100-103.
- Cullum, C. M., Heaton, R. K., & Grant, I. (1991). Psychogenic factors influencing neuropsychological performance: Somatoform disorders, factitious disorders, and malingering. In H. O. Doerr & A. Carlin (Eds.), *Forensic neuropsychology* (pp. 141-171). New York: Guilford Press.
- Cullum, C. M., & Rosenberg, R. N. (1998). Memory loss—when is it Alzheimer disease? *Journal of the American Medical Association*, 279, 1689-1690.
- Cullum, C. M., Thompson, L. L., & Smernoff, E. N. (1993). Three word recall as a measure of memory. *Journal of Clinical and Experimental Neuropsychology*, 15, 321-329.
- Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California verbal learning test*. Research edition. San Antonio, TX: The Psychological Corporation.
- Delis, D. C., Massman, P. J., Kaplan, E., McKee, R., Kramer, J. H., & Gettman, D. (1991). Alternate form of the California Verbal Learning Test: Development and reliability. *The Clinical Neuropsychologist*, 5, 154-162.
- Delis, D. C., Squire, L. R., Bihler, A., & Massman, P. (1992). Componential analysis of problem-solving ability: Performance of patients with frontal lobe damage and amnesic patients on a new sorting test. *Neuropsychologia*, 30, 683-697.
- Dodrill, C. B. (1997). Myths of neuropsychology. *The Clinical Neuropsychologist*, 11, 1-17.
- Eubanks, J. D. (1997). Clinical neuropsychology summary information prepared by Division 40, Clinical Neuropsychology, American Psychological Association. *The Clinical Neuropsychologist*, 11, 77-80.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." *Journal of Psychiatric Research*, 12, 189-198.
- Franzen, M. D., Burgess, E. J., & Smith-Seemiller, L. (1998). Methods of estimating premorbid functioning. *Archives of Clinical Neuropsychology*, 12, 711-738.
- Franzen, M. D., Iverson, G. L., & McCracken, L. M. (1990). The detection of malingering in neuropsychological assessment. *Neuropsychology Review*, 1, 247-279.
- Freedman, M., Leach, L., & Kaplan, E. (1994). *Clock drawing: A neuropsychological analysis*. New York: Oxford University Press.
- Gauthier, L., Dehaut, F., & Joannette, Y. (1989). The Bells Test: A quantitative and qualitative test for visual neglect. *International Journal of Clinical Neuropsychology*, 11, 49-54.
- Goodglass, H., & Kaplan, E. (1987). *The assessment of aphasia and related disorders* (2nd ed.). Philadelphia: Lea & Febiger.
- Grant, I., & Adams, K. M. (Eds.) (1996). *Neuropsychological Assessment of neuropsychiatric disorders* (2nd ed.). New York: Oxford University Press.
- Greve, K. W., Farrwell, J. F., Besson, P. S., & Crouch, J. A. (1995). A psychometric analysis of the California Card Sorting Test. *Archives of Clinical Neuropsychology*, 10, 265-278.
- Guilmette, T., Hart, K., & Giuliano, A. (1993). Malingering detection: The use of a forced choice method in identifying organic versus simulated memory impairment. *The Clinical Neuropsychologist*, 7, 59-69.
- Halstead, W. C. (1947). *Brain and intelligence*. Chicago: University of Chicago Press.

- Hannay, H. J., Bieliauskas, L. A., Crosson, B. A., Hammeke, T. A., Hamsher, K. De-S., & Koffler, S. P. (1998). Proceedings of the Houston Conference on Specialty Education and Training in Clinical Neuropsychology. *Archives of Clinical Neuropsychology*, 13, 157-250.
- Heaton, R. K. (1981). *Wisconsin Card Sorting Test (WCST)*. Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K. (1992). *Comprehensive norms for an expanded Halstead-Reitan battery: A supplement for the Wechsler Memory Scale-Revised*. Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtis, G. (1993). *Wisconsin Card Sorting Test (WCST) Manual Revised and expanded*. Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan battery: Demographic corrections, research findings, and clinical applications*. Odessa, FL: Psychological Assessment Resources.
- Heaton, R. K., Ryan, L., Grant, I., & Matthews, C. G. (1996). Demographic influences on neuropsychological test performance. In K. M. Adams & I. Grant (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 141-163). New York: Oxford University Press.
- Hermann, B. P., Wyler, A. R., Richey, E. T., & Rea, J. M. (1987). Memory function and verbal learning ability in patients with complex partial seizures of temporal lobe origin. *Epilepsia*, 28, 547-554.
- Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology*, 11, 967-974.
- Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokman, E., & Kurland, L. T. (1992). Mayo's older Americans normative studies: WMS-R norms for ages 56 to 94. *The Clinical Neuropsychologist*, 6, 49-82.
- Kaplan, E. (1990). The process approach to neuropsychological assessment of psychiatric patients. *Journal of Neuropsychiatry*, 2, 72-87.
- Kaplan, E., Fein, D., Morris, R., & Delis, D. C. (1991). *WAIS-R as a neuropsychological instrument*. San Antonio, TX: The Psychological Corporation.
- Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test* (2nd ed.). Philadelphia: Lea & Febiger.
- Kertesz, A. (1982). *Western Aphasia Battery*. San Antonio, TX: The Psychological Corporation.
- Klove, H. (1963). *Grooved pegboard*. Lafayette, IN: Lafayette Instruments.
- Kozora, E., & Cullum, C. M. (1994). Qualitative features of clock drawings in normal aging and Alzheimer's disease. *Assessment*, 1, 179-187.
- Kozora, E., & Cullum, C. M. (1995). Generative naming in normal aging: Total output and qualitative changes using phonemic and semantic constraints. *The Clinical Neuropsychologist*, 9, 313-320.
- Lacritz, L. H., & Cullum, C. M. (1998). Stroke and depression. In A. J. Rush (Ed.), *Mood & anxiety disorders* (pp. 339-362). Philadelphia: Williams and Wilkins.
- Lansing, A. E., Ivnik, R. J., Cullum, C. M., & Randolph, C. (in press). An empirically derived short form of the Boston Naming Test. *Archives of Clinical Neuropsychology*.
- Lewis, R., & Kupke, T. (1977). *The Lafayette Clinic Repeatable Neuropsychological Test Battery: Its development and research applications*. Paper presented at the annual meeting of the Southeastern Psychological Association, Hollywood, Florida.
- Lewis, R. F., & Rennick, P. M. (1979). *Manual for the Repeatable Cognitive-Perceptual-Motor Battery*. Clinton Township, MI: Ronald F. Lewis.
- Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.
- Libon, D. J., Malamut, B. L., Swenson, R., Sands, L. P., & Cloud, B. S. (1996). Further analyses of clock drawings among demented and nondemented older subjects. *Archives of Clinical Neuropsychology*, 11, 193-205.
- Libon, D. J., Mattson, R. E., Glosser, G., Kaplan, E., Malamut, B. L., Sands, L. P., Swenson, R., & Cloud, B. S. (1996). A nine-word dementia version of the California Verbal Learning Test. *The Clinical Neuropsychologist*, 10, 237-244.
- Loring, D. W., Lee, G. P., Martin, R. C., & Meador, K. J. (1989). Verbal and visual memory index discrepancies from the Wechsler Memory Scale-Revised: Cautions in interpretation. *Psychological Assessment*, 1, 198-202.
- Loring, D. W., Martin, R. C., Meador, K. J., & Lee, G. P. (1990). Psychometric construction of the Rey-Osterrieth complex figure: Methodological considerations and interrater reliability. *Archives of Clinical Neuropsychology*, 5, 1-14.
- Luria, A. R. (1966). *The working brain: An introduction to neuropsychology*. New York: Basic Books.
- Malloy, P. F., & Richardson, E. D. (1998). Assessment of frontal lobe functions. *Journal of Neuropsychiatry*, 6, 399-410.
- Markwardt, F. C., Jr. (1989). *The Peabody Individual Achievement Test-Revised*. Circle Pines, MN: American Guidance Service.
- Matarazzo, J. D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic, and courtroom. *American Psychologist*, 45, 999-1017.
- Matarazzo, J. D., & Prifitera, A. (1989). Subtest scatter and premorbid intelligence: Lessons from the WAIS-R standardization sample. *Psychological Assessment*, 1, 186-191.
- Mattis, S. (1988). *Dementia Rating Scale (DRS)*. Odessa, FL: Psychological Assessment Resources.
- Matt-Maddrey, A., Cullum, C. M., Weiner, M. F., & Filley, C. M. (1996). Premorbid intelligence estimation and level of dementia in Alzheimer's disease. *Journal of The International Neuropsychological Society*, 2, 1-5.
- Meyers, J. E., & Meyers, K. R. (1995). Rey complex figure test under four different administration procedures. *The Clinical Neuropsychologist*, 9, 63-67.
- Milberg, W. P., Hebben, N., & Kaplan, E. (1996). The Boston process approach to neuropsychological assessment. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (2nd ed., pp. 58-80). New York: Oxford University Press.
- Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review*, 2, 109-145.
- Morris, J. C., Heyman, A., Mohs, R. C., et al. (1989). The consortium to establish a registry for Alzheimer's disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, 39, 1159-1165.
- Moscovitch, M. (1994). Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. *Neuropsychology*, 8, 524-534.
- Naugle, R. I., Chelune, G. J., Cheek, R., Luders, H., & Awad, I. A. (1993). Detection of changes in material-specific memory following temporal lobectomy using the Wechsler Memory Scale-Revised. *Archives of Clinical Neuropsychology*, 8, 381-395.
- Naugle, R. I., Chelune, G. J., Schuster, J., Luders, H. O., & Comair, Y. (1994). Recognition memory for words and

- faces before and after temporal lobectomy. *Assessment*, 1, 373–381.
- Naugle, R. I., Cullum, C. M., & Bigler, E. D. (1998). *Introduction to clinical neuropsychology: A casebook*. Austin, TX: ProEd.
- Naugle, R. I., Cullum, C. M., Bigler, E. D., & Massman, P. J. (1986). Neuropsychological and computerized axial tomography volume characteristics in senile and pre-senile dementia. *Archives of Clinical Neuropsychology*, 1, 219–230.
- Nelson, H. E. (1982). *The National Adult Reading Test (NART): Test Manual*. Windsor, UK: NFER-Nelson.
- Nelson, H. E., & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the National Adult Reading Test. *Cortex*, 14, 234–244.
- Nies, K. J., & Sweet, J. J. (1994). Neuropsychological assessment and malingering: A critical review of past and present strategies. *Archives of Clinical Neuropsychology*, 9, 501–552.
- Osterrieth, P. A. (1944). Le test de copie d'une figure complexe. *Archives de Psychologie*, 30, 206–356; translated by J. Corwin & F.W. Bylsma (1993), *The Clinical Neuropsychologist*, 7, 9–15.
- Paolo, A. M., Tröster, A. I., & Ryan, J. J. (1997). Test-retest stability of the California Verbal Learning Test in older persons. *Neuropsychology*, 11, 613–616.
- Peavy, G., Jacobs, D., Salmon, D., Butters, N., Delis, D., Taylor, M., Massman, P., Stout, J., Heindel, W., Kirson, E., Atkinson, J., Chandler, J., Grant, I., and the HIV Neurobehavioral Research Center Group (1994). Verbal memory performance of patients with human immunodeficiency infection: Evidence of subcortical dysfunction. *Journal of Clinical and Experimental Neuropsychology*, 16, 508–523.
- Powell, D. H., Kaplan, E. F., Whitla, D., Weintraub, S., Catlin, R., & Funkenstein, H. H. (1993). *Manual for MicroCog: Assessment of cognitive functioning*. San Antonio, TX: The Psychological Corporation.
- Rapport, L. J., Dutra, R. L., Webster, J. S., Charter, R., & Morrill, B. (1995). Hemispatial deficits on the Rey-Osterrieth complex figure drawing. *The Clinical Neuropsychologist*, 9, 169–179.
- Raven, J. C. (1994). *Advanced progressive matrices sets I & II*. Oxford, UK: Oxford Psychologists Press.
- Raven, J. C. (1995). *Colored progressive matrices sets A, Ab, B*. Oxford, UK: Oxford Psychologists Press.
- Raven, J. C. (1996). *Progressive Matrices: A perceptual test of intelligence. Individual Form*. Oxford, UK: Oxford Psychologists Press.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.
- Reitan, R. M., & Wolfson, D. (1996). Relationships of age and education to Wechsler Adult Intelligence Scale IQ values in brain-damaged and non-brain-damaged groups. *The Clinical Neuropsychologist*, 10, 293–304.
- Reports of the INS-Division 40 Task Force on Education, Accreditation, and Credentialing (1987). *The Clinical Neuropsychologist*, 1, 29–34.
- Reports of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology (1996). Assessment: Neuropsychological testing of adults. *Neurology*, 47, 592–599.
- Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*, 28, 286–340.
- Reynolds, C. R. (1997). Forward and backward memory span should not be combined for clinical analysis. *Archives of Clinical Neuropsychology*, 12, 29–40.
- Ross, E. D. (1985). Modulation of affect and nonverbal communication by the right hemisphere. In M. M. Mesulam (Ed.), *Principles of behavioral neurology* (pp. 239–257). Philadelphia: F.A. Davis.
- Royall, D. R., Cordes, J. A., & Polk, M. (1998). CLOX: An executive clock drawing task. *Journal of Neurology, Neurosurgery, and Psychiatry*, 64, 588–594.
- Russell, E. W. (1975). A multiple-scoring method for the assessment of complex memory functions. *Journal of Consulting and Clinical Psychology*, 43, 800–809.
- Russell, E.W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach*. New York: Wiley.
- Ryan, C. M., Morrow, L. A., Bromet, E., & Parkinson, D. K. (1987). Assessment of neuropsychological dysfunction in the workplace: Normative data from the Pittsburgh occupational exposures test battery. *Journal of Clinical and Experimental Neuropsychology*, 9, 665–679.
- Sass, K. J., Sass, A., Westerveld, M., Lencz, T., Novelly, R. A., Kim, J. H., & Spencer, D. D. (1992). Specificity in the correlation of verbal memory and hippocampal neuron loss: Dissociation of memory, language, and verbal intellectual ability. *Journal of Clinical and Experimental Neuropsychology*, 14, 662–672.
- Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 1–38). Cambridge, MA: MIT Press.
- Silverstein, A. B. (1982). Two-and four-subtest short forms of the Wechsler Adult Intelligence Scale-Revised. *Journal of Consulting and Clinical Psychology*, 50, 415–418.
- Smith, G. E., Ivnik, R. J., Malec, J. F., Kokmen, T., Tangalos, E., & Petersen, R. C. (1994). Psychometric properties of the Mattis Dementia Rating Scale. *Assessment*, 1, 123–131.
- Spreen, O., & Strauss, E. (1998). *Compendium of neuropsychological tests* (2nd ed.). New York: Oxford University Press.
- Squire, L. R. (1987). *Memory and brain*. New York: Oxford University Press.
- Stern, R. A., Singer, E. A., Duke, L. M., Singer, N. G., Morey, C. E., Daughtrey, E. W., & Kaplan, E. (1994). The Boston qualitative scoring system for the Rey-Osterrieth Complex Figure: Description and interrater reliability. *The Clinical Neuropsychologist*, 8, 309–322.
- Sternberg, R. J. (Ed.) (1997). Intelligence and lifelong learning [Special issue]. *American Psychologist*, 52.
- Strub, R. L., & Black, F. W. (1988). *Neurobehavioral disorders: A clinical approach*. Philadelphia: F.A. Davis.
- Stuss, D. T., & Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin*, 95, 3–28.
- Sweet, J. J., & Moberg, P. J. (1990). A survey of practices and beliefs among ABPP and non-ABPP clinical neuropsychologists. *The Clinical Neuropsychologist*, 4, 101–120.
- Tombaugh, T. N., McDowell, I., Krisjansson, B., & Hubley, A. M. (1996). Mini-Mental State Examination (MMSE) and the Modified MMSE (3MS): A psychometric comparison and normative data. *Psychological Assessment*, 8, 48–59.
- Trenerry, M. R., Crosson, B., DeBoe, J., & Leber, W. R. (1990). *Visual Search and Attention Test*. Odessa, FL: Psychological Assessment Resources.
- Tröster, A. I., Jacobs, D., Butters, N., Cullum, C. M., & Salmon, D. P. (1989). Differentiating Alzheimer's disease from huntington's disease with the Wechsler Memory Scale-Revised. *Clinics in Geriatric Medicine*, 5, 611–632.
- Warrington, E. K. (1984). *Recognition Memory Test*. Windsor, UK: NFER-Nelson.
- Wechsler, D. (1945). A standardized memory scale for clinical use. *The Journal of Psychology*, 19, 87–95.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1987). *Wechsler Memory Scale-Revised*. San Antonio, TX: The Psychological Corporation.

- Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale-Third Edition*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997b). *Wechsler Memory Scale-Third Edition*. San Antonio, TX: The Psychological Corporation.
- Welsh, K. A., Butters, N., Mohs, R. C., Beekly, D., Edland, S., Fillenbaum, G., & Heyman, A. (1994). The consortium to establish a registry for Alzheimer's disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology*, 44, 609–614.
- Wilkinson, G. S. (1993). *The Wide Range Achievement Test-3 manual*. Wilmington, DE: Wide Range, Inc.
- Woodcock, R. W., & Mather, N. (1989). *Woodcock-Johnson Tests of Achievement*. Allen, TX: DLM Teaching Resources.

4.12

Principles of Personality Assessment

JERRY S. WIGGINS and KRISTA K. TROBST
University of British Columbia, Vancouver, BC, Canada

4.12.1 INTRODUCTION	350
4.12.2 PSYCHODYNAMIC TRADITION	350
4.12.2.1 <i>Historical Background</i>	350
4.12.2.2 <i>Conceptual Framework</i>	351
4.12.2.3 <i>Assessment Instruments</i>	351
4.12.2.4 <i>Interpretive Principles</i>	351
4.12.2.4.1 <i>Projective hypothesis</i>	352
4.12.2.4.2 <i>Levels of functioning</i>	352
4.12.2.4.3 <i>Psychological adjustment</i>	352
4.12.2.5 <i>Applications and Current Status</i>	352
4.12.3 PERSONOLOGICAL TRADITION	353
4.12.3.1 <i>Historical Background</i>	353
4.12.3.2 <i>Conceptual Framework</i>	353
4.12.3.3 <i>Assessment Instruments</i>	354
4.12.3.4 <i>Interpretive Principles</i>	354
4.12.3.5 <i>Applications and Current Status</i>	355
4.12.4 MULTIVARIATE TRADITION	355
4.12.4.1 <i>Historical Background</i>	355
4.12.4.2 <i>Conceptual Framework</i>	356
4.12.4.3 <i>Assessment Instruments</i>	356
4.12.4.4 <i>Interpretive Principles</i>	356
4.12.4.5 <i>Applications and Current Status</i>	357
4.12.5 EMPIRICAL TRADITION	357
4.12.5.1 <i>Historical Background</i>	357
4.12.5.2 <i>Conceptual Framework</i>	358
4.12.5.3 <i>Assessment Instruments</i>	359
4.12.5.4 <i>Interpretive Principles</i>	359
4.12.5.4.1 <i>Empirical strategy</i>	359
4.12.5.4.2 <i>Constructive strategy</i>	359
4.12.5.4.3 <i>Interpersonal strategy</i>	359
4.12.5.5 <i>Applications and Current Status</i>	360
4.12.6 INTERPERSONAL PARADIGM	360
4.12.6.1 <i>Historical Background</i>	360
4.12.6.2 <i>Conceptual Framework</i>	361
4.12.6.3 <i>Assessment Instruments</i>	361
4.12.6.3.1 <i>Interpersonal Adjective Scales</i>	361
4.12.6.3.2 <i>Inventory of Interpersonal Problems</i>	361
4.12.6.3.3 <i>Impact Message Inventory</i>	362
4.12.6.4 <i>Interpretive Principles</i>	362
4.12.6.5 <i>Applications and Current Status</i>	363

4.12.1 INTRODUCTION

Consider a hypothetical clinician who is both well trained and well informed about available personality tests. In her daily practice, she employs a range of assessment instruments (singly or in combination) to address the range of diagnostic issues posed by her clients. When “thought disorder” is suspected, she might occasionally administer a Rorschach and from most clients she will obtain a case history. Her clients themselves may be asked to fill out the NEO-Personality Inventory to clarify dimensions of normal personality functioning or the Minnesota Multiphasic Personality Inventory-Revised (MMPI-2) to assess gross abnormality. She might also routinely administer the Interpersonal Adjective Scales to assess the character and quality of clients’ interpersonal relationships.

Like many of her colleagues, our clinician may describe her theoretical orientation as “eclectic” and indeed that is what she appears to be. But she is also, perhaps unknowingly, operating within larger historical contexts of assumptions and beliefs that may include traditions of thought as diverse as psychoanalysis, psychobiography, psychometrics, empirical realism, and symbolic interactionism (respectively for the tests just mentioned). We will refer to such contexts as different “traditions” of personality assessment, by which we mean the background of a set of unquestioned beliefs or orienting attitudes within and against which personality tests are administered and interpreted.

The traditions that will be considered in this chapter are referred to as psychodynamic, personological, multivariate, empirical, and interpersonal. Such traditions are related, but *not* equivalent, to certain theories of personality (Hall & Lindzey, 1978) and psychopathology (Millon, 1969). Similarly, these traditions are related, but *not* equivalent, to certain kinds of personality tests. Thus, the MMPI has been associated with a distinctive philosophy of science (empirical realism) rather than with a particular theory of psychopathology. The Rorschach test and the Wechsler scales have figured prominently within the ego psychological perspective of the psychodynamic tradition, although both tests were constructed for quite different purposes.

We believe that these traditions reflect different and potentially useful ways of viewing the entire psychodiagnostic enterprise. We also

believe that these traditions have contributed to more or less invisible barriers to communication being erected among clinicians who work, knowingly or not, within them. This state of affairs is unfortunate because “membership” in a tradition, as it were, is mainly determined by where one went to graduate school, with whom one studied, and the particular clinical setting in which one finds oneself operating. Further, we are convinced that, at the present time, there are some remarkably similar ideas being advocated about the nature of personality and its assessment among those working within several different traditions, and that recognition of these commonalities might advance the field as a whole (see Wiggins, in press).

We would venture to suggest that most readers will have familiarity with a few of these traditions, but not with some of the others. For example, those intimately familiar with the psychodynamic tradition are less likely to be familiar with the multivariate and empirical traditions, and vice versa. Those who are fond of obtaining life history data from interviews may not be fully aware of the rich history of the personological tradition nor of the exciting recent developments that have occurred within that tradition. And many may not be aware that the interpersonal tradition has fostered a strong “underground movement” within clinical psychology for many decades.

4.12.2 PSYCHODYNAMIC TRADITION

4.12.2.1 Historical Background

David Rapaport (1911–1960) was a major systematist of psychoanalytic theory, the originator of a now standard psychodiagnostic test battery, and a highly influential mentor of the principal architects of the psychodynamic tradition in personality assessment. After receiving his PhD from the Royal Hungarian University, he emigrated to the USA in 1938 and shortly thereafter joined the staff of the Menninger Clinic where he eventually became chief psychologist and head of the research department. During and shortly after World War II, the results of an extensive program of collaborative research were summarized in a two-volume *Manual of diagnostic psychological testing* (Rapaport, 1944–46) that eventually “revolutionized clinical psychology and influenced clinical psychologists the world over” (Gill & Klein, 1967, p. 18).

Rapaport moved to the Austin Riggs Center in 1948 and continued his extensive collaborations with colleagues at institutions such as the Menninger Foundation, Yale University Department of Psychiatry, and the Research Center for Mental Health at New York University. Speaking collectively for workers at these and other institutions, Roy Schafer (1967) observed, "All of us are working within the psychoanalytic psychodiagnostic tradition crystallized by David Rapaport" (p. 2).

4.12.2.2 Conceptual Framework

It is difficult to make a coherent set of distinctions among the many "psychodynamic" theories that have been proposed as alternatives to Freud's classical psychoanalytic theory (Westen, 1990). It would seem to be generally true however that since, and even before, Freud's death, there has been an increased emphasis on the autonomous and adaptive functions of the ego and on the internal representation of self and others within the ego domain. The aspect of classical psychoanalysis that has been de-emphasized, or even abandoned by some, is the proposition that personality and its development are largely determined by unconscious libidinal and aggressive drives.

Rapaport (1960) must be counted among the most rigorous systematists of classical psychoanalytic theory and he was also a major contributor to the alternative perspective known as "ego psychology." In both his theoretical and assessment work, his principal interest was in thought processes and their course of development from drive-oriented to reality-oriented processes. In this context, Hartmann's (1939) concept of the independence (from drive) of the ego and its role in fostering active processes of adaptation were appealing to him, as was Erikson's (1950) emphasis on the cultural and historical contexts in which the ego develops. Rapaport (1960) developed and expanded the metaconcept of structure which he felt was the "prime requisite for progress toward dimensional quantification" (p. 98) and which he applied broadly to many aspects of psychological functioning and cognitive organization (see Gill & Klein, 1967).

The generally increased emphasis upon ego functions in postclassical psychoanalysis was also reflected in the currently important psychodynamic alternative of object relations theory. Proponents of this view challenged the classical idea that cathexes of external "objects" (including persons) served mainly as vehicles through which instinctual energies were discharged. Instead, it was postulated that early

interactions with significant others ("objects") led to internalized representations of both others (Jacobson, 1954) and self (Kohut, 1971) that serve as "internal working models" (Bowlby, 1973) for later interpersonal relationships. From this perspective, the person is, from birth (Main, Kaplan, & Cassidy, 1985), an object seeker (Fairbairn, 1952) who establishes a "gratifying involvement" with other persons (Behrends & Blatt, 1985).

The shift in emphasis within psychoanalytic theory from drives to internalized representations has been accompanied by a corresponding shift in rationales for personality assessment, as exemplified by the ego-psychological approach of Allison, Blatt, and Zimet (1968) and the object-relations approach of Blatt and Lerner (1983).

4.12.2.3 Assessment Instruments

Prior to Rapaport's influential contributions in the 1940s, assessment psychologists were primarily technicians who administered IQ tests. Since that time, they have become clinicians who administer batteries of both structured and projective tests of personality and cognition. The multitest battery was advocated in view of the apparent complexity of personality and cognition and their inter-related functions, as well as for the purpose of gathering normative data that would shed light on those complexities (Rapaport, Gill, & Schafer, 1946). At Menninger, the composition of the battery reflected judgments regarding the potential of each instrument to yield measures that might be interpreted within the ego psychological framework of Rapaport and his associates.

The projective component of the original Menninger battery included the Rorschach Test (Rorschach, 1921), the Thematic Apperception Test (TAT; Morgan & Murray, 1935), and a locally constructed Word Association Test. The nonprojective component included the Bellevue Scale (Wechsler, 1941), the Babcock Test of mental efficiency (Babcock, 1933), the Sorting Test of concept formation (Goldstein & Scheerer, 1941), and the Hanfmann-Kasanin test of concept formation (Hanfmann & Kasanin, 1937). As this test battery evolved over a 20-year period, the Rorschach, TAT, and WAIS (Wechsler, 1958) became the more or less standard core of the psychodynamic test battery (e.g., Allison et al., 1968).

4.12.2.4 Interpretive Principles

The vast literature on interpretive principles associated with the psychodynamic paradigm

defies easy summarization. Rapaport's theoretical work (e.g., Gill, 1967) and the writings of those whom he influenced (e.g., Holt, 1967) are, understandably, highly technical in nature. Of greater help to the clinician are the case summaries (Allison et al., 1968; Prelinger & Zimet, 1964; Rapaport et al., 1946; Schafer, 1948) and the individual treatments of the Rorschach (Schafer, 1954), TAT (Holt, 1951), and WAIS (Blatt & Allison, 1981). The following statements regarding the projective hypothesis, levels of functioning, and psychological adjustment are meant only to convey some of the flavor of three, among many, important interpretive principles.

4.12.2.4.1 Projective hypothesis

The projective hypothesis states that "All behavior manifestations of the human being, including the least and the most significant, are revealing and expressive of his personality, by which we mean that individual principle of which he is the carrier" (Rapaport, 1942, p. 92). Thus, people's possessions—clothes, automobiles, furniture—are expressive of their personalities and reflect single acts of choice; in their totality, they reflect the organization of such choices. Responses to the ambiguous stimuli of projective tests may also be thought of in terms of "choice," although such choices are much less conscious or volitional in nature. Thus, responses to a Rorschach inkblot may be thought of as reflecting "choices" between forms, colors, shadings, and so forth, to which a subject imparts meaning through organization. Responses to a TAT card also involve both choice (e.g., which figure to identify with) and organization (e.g., sequence of events). Responses to intelligence and concept formation tests involving choice and organizational processes may be used as "nonprojective tests of personality," given an adequate theory of "functions underlying the reactions and achievements on these tests" (Rapaport, 1946, p. 228).

4.12.2.4.2 Levels of functioning

The psychoanalytic model of primary (pleasure principle) and secondary (reality principle) modes of thought is meant to account for both a developmental sequence and characteristics of the mature adult (Rapaport, 1951). Consequently, there is a continuum of adult psychological functioning that may be assessed with an appropriate battery of tests.

This continuum ranges from functioning in situations which put a premium on highly logical,

reality-oriented secondary modes of thought (WAIS) to those which allow for more personal, less conventionally constrained thinking (TAT) and finally those which allow for considerably novel, personalized, and regressive modes of thinking (Rorschach). (Allison et al., 1968, p. vii)

Assessment of level of functioning has been greatly facilitated by Holt's innovative procedures for assessing primary and secondary process in the Rorschach (Holt & Havel, 1960).

4.12.2.4.3 Psychological adjustment

The Rorschach, TAT, and WAIS may be employed to assess both adaptive capacities and impairments in psychological functioning and to identify the functions impaired in different psychiatric diagnostic groups. Adjustment and maladjustment may be assessed with reference to the following postulated sequence:

Certain patterns of defense mechanisms are adopted and these determine specific strengths and weaknesses in psychological functioning which then become characteristic of the adjustment of the personality; with the onset of maladjustment, an exaggeration or breakdown in these strengths and weaknesses characteristic for that maladjustment occurs which can be measured; this leads to a diagnostic differentiation. (Rapaport, Menninger, & Schafer 1947, p. 249)

4.12.2.5 Applications and Current Status

Since the late 1960s, theory and method within the psychodynamic paradigm of personality assessment have evolved into an object relations perspective that is highly compatible with contemporary formulations of social cognition, information processing, attachment research, and ego development (see Westen, 1990). Sidney Blatt and his associates at Yale have been primarily responsible for this paradigm shift. From the early 1950s until his untimely death in 1960, Rapaport contributed to the development of ego psychology and its application to personality assessment; since the early 1970s, Blatt has contributed to the development of object relations theory and its application to current personality assessment methods.

"The study of the representation of the human form on the Rorschach is an ideal data base for assessing an individual's representational world—his conception of people, including himself, and their actual and potential interactions. The representation of people, that is, object representations, have both structure and content" (Blatt & Lerner, 1983, p. 8). The structural aspects of object representations are

emphasized in the Rorschach scoring system developed by Blatt, Brenneis, Schimek, and Glick (1976) that provides a developmental analysis of object representations in terms of such categories as differentiation, articulation, and integration of object and action. The content and affective themes of object representations are emphasized in the Rorschach scoring system developed by Mayman (1967) that emphasizes phenomenological dimensions such as affect states, ego states, experience of self, and sense of identity.

An object relations scoring system for the TAT has also been developed by Westen (1991) in which stories are rated for complexity of representations of people, affective tone of relationship paradigms, capacity for emotional investment, and understanding of social causality. A similar scoring system has been developed for the Picture Arrangement subtest of the WAIS (Westen, 1991, p. 72). In conclusion, the three major instruments of the original Menninger battery continue to show promise under a revised conceptual orientation that is most compatible with current thinking in personality, social, clinical, and developmental psychology.

4.12.3 PERSONOLOGICAL TRADITION

4.12.3.1 Historical Background

The “storied” nature of human conduct (Sarbin, 1986) has been recognized for many centuries within the disciplines of history, literature, and most of the social sciences. Within the much more recent personological tradition in personality assessment, the person is taken to be the basic unit of observation and the person’s life story is considered to be the preferred way in which the person is to be understood (McAdams, 1994). The origins of the personological tradition, and in fact the origins of academic personality psychology itself, can be traced most directly to Harvard University in the 1930s. At that institution, Gordon Allport (1937) wrote the first textbook on personality, which defined the person as the basic unit of observation. Subsequently, he emphasized the use of personal documents in psychobiography (Allport, 1942, 1965) and spent a lifetime pondering the question of “How shall a psychological life history be written?” (Allport, 1967). At the Harvard Psychological Clinic, Murray (1938) introduced a multiform organismic method for studying “lives in progress” (White, 1952), in which an interdisciplinary team of investigators applied a variety of assessment procedures to the study of individuals over the course of their lives. During

World War II, a variant of this procedure was employed in the selection of intelligence agents and saboteurs (Office of Strategic Services, 1948), and subsequently this approach spawned a variety of successful peacetime assessment programs (e.g., MacKinnon, 1975; Stern, Stein, & Bloom, 1956).

Murray’s enduring contributions to the personological paradigm go well beyond those just mentioned. During his 60 years at Harvard, Murray was a source of personal inspiration for an extraordinarily talented and diverse group of students, colleagues, and visitors who, over the years, have applied and expanded his “personology” up until the present time. Early associates at the clinic included Samuel Beck, Erik Erikson, Jerome Frank, Daniel Levenson, Donald MacKinnon, Silvan Tomkins, and Robert White. Dan McAdams and William McKinley Runyan are more recent products of this Harvard tradition.

4.12.3.2 Conceptual Framework

Murray’s (1938) personology provided a set of orienting attitudes that guided the multiform (many assessors, many tests), organismic (holistic) approach to the study of lives over time (e.g., White, 1966, 1975). Murray’s (1959) conceptual framework was meant to serve as “the scaffold of a comprehensive system,” rather than as a completed formal theory. Nevertheless, Murray (1938) introduced a number of concepts that have subsequently proven useful in guiding the study of lives. By taking the life cycle of the individual as the largest unit of study, Murray was committed to studying personality “the long way” (White, 1981), and he introduced units of time varying in duration and complexity from the proceeding (single episode) to the unity-thema (central life motif). A thema may be understood in terms of the interaction of needs (for which Murray developed his famous taxonomy) and press (environmental facilitation or obstruction of a need). The TAT (Murray, 1943) was developed as a projective measure of need–press interaction and has been utilized in large-scale studies of needs for achievement (McClelland, 1961), power (Winter, 1973), and intimacy (McAdams, 1989).

Erik Erikson’s (1950) psychodynamic theory of personality development provided a more explicit account of the stages of psychosocial development over the entire life span (from the original psychosocial crisis of “trust versus mistrust” to the final crisis of “ego-integrity versus despair”). His theory informed his own classic psychobiographies, such as those of

Martin Luther (Erikson, 1958) and Gandhi (Erikson, 1969), as well as contributing to more recent psychobiographical efforts (e.g., Stewart, Franz, & Layton, 1988). More recently, in a book entitled *The stories we live by: Personal myths and the making of the self*, McAdams (1993) presented a neo-Eriksonian theory of identity development that is enriched by more contemporary concepts such as image, prototype, and script, and that is guided by the metatheoretical concepts of agency and communion (Bakan, 1966). This work has been hailed as “the most original and important new book on personality theory since George Kelly’s *Psychology of personal constructs*” (Hogan, 1994, p. 356). The heuristic potential of McAdams’ conceptual framework is suggested by the quality and quantity of critical response it has elicited from personality psychologists representing diverse theoretical perspectives (see McAdams, 1996 and peer commentary therein). Although McAdams’ framework is more formal, explicit, and testable than earlier efforts, its roots in the formulations of Allport and Murray are easily traceable (McAdams, 1994).

4.12.3.3 Assessment Instruments

In what is now widely recognized as the definitive general treatise on life histories and psychobiography, Runyan (1982) observed:

It should be clear that there is no single life history method, any more than there is a single personality research method, and that life histories may be studied through phenomenological self-reports, archival research, prospective longitudinal research, and experimental research. (p. 6)

Personality assessment psychologists from Murray to McAdams have recognized the need for methodological pluralism (Craik, 1986) in the study of lives, and for that reason the “boundaries” of the personological paradigm have historically been somewhat more permeable than those of the other paradigms considered in this chapter. Nevertheless, it would be inaccurate to characterize this paradigm as “eclectic,” because the diversity of methods employed reflects the different levels of analysis encompassed by the term “life history,” rather than the selective use of theories and methods to “explain” a given life. Following Kluckhohn and Murray (1953), Runyan (1982) distinguished three levels of generality in the social sciences: the universal (general laws of human behavior); the group (differences in sex, social class, culture); and the individual (distinctive, distinguishing characteristics of a par-

ticular person) and surveyed attempts to establish acceptable criteria for interpretations at each level. Although personologists in general, and clinicians in particular, are primarily concerned with the individual, an awareness of the larger enterprise that Runyan portrays serves to clarify the nature and goals of individual personality assessment. And for those few who might seriously be considering becoming psychobiographers, Alan Elms’ (1994) book is essential reading.

The 25 different assessment procedures employed for the study of individuals at the Harvard Clinic (Murray, 1938) were indeed “multiform” in terms of both methods and the assessors who employed them; for example, autobiography (Murray), hypnosis (White), level of aspiration (Frank), dramatic productions (Erikson), and Rorschach (Beck). Perhaps of greatest interest to the contemporary clinician is the case study method in which some form of interview is employed with all the attendant advantages and limitations of retrospective, introspective, and qualitative methods (see Runyan, 1982, chap. 8). The procedures employed by McAdams (1993, chap. 10) provide an excellent example of a theory-driven approach to the interview.

4.12.3.4 Interpretive Principles

The variety of personality research methods for studying an individual’s life history is reflected in a similar variety of interpretive principles associated with different methods. In this context, the interpretive principles elucidated by Alexander (1990) would appear to be “mainstream” in the sense that they evolved directly from the framework of Murray and his co-workers (especially Tomkins 1947, 1979) at the Harvard Psychological Clinic.

In analyzing data from directed interviews or autobiographical essays, Alexander (1990) looks for recurring dynamic sequences (scripts, themas, guiding messages) that may be revealed by “letting the data speak” or by “asking the data a question.” In the former method, the typically large data set may be reduced by identifying “significant” material according to the following nine criteria of salience: (i) primacy (that which occurs first), (ii) frequency (that which occurs often), (iii) uniqueness (that which is unusual or odd), (iv) negation (that which is denied or disavowed), (v) emphasis (that which is either overemphasized or underemphasized), (vi) omission (that which is missing by normative standards or by implication), (vii) error or distortion (factual errors and Freudian slips), (viii) isolation (that which does

not “fit” and non sequiturs), and (ix) incompleteness (that which is not finished or lacks closure).

The method of “asking the data a question” is similar to that employed in the analysis of TAT protocols for predetermined categories such as “power motivation” (Winter, 1973). The database of interview material is reduced by selecting every sequence or incident related to the questions posed by the investigator. When applied to interview material, this method retains the “projective” advantage of addressing issues that the storyteller did not consciously intend to describe or deal with, while circumventing problems such as the identification of the “hero” of the narrative (Alexander, 1990).

4.12.3.5 Applications and Current Status

In recent years, there appears to have been a “back to basics” movement in personality assessment research (Wiggins & Pincus, 1992) that has rekindled interest in the fundamental assumptions of several paradigms, including the personological: “Once, again, it is okay to study the ‘whole person.’ Better, contemporary personologists insist, as did pioneers like Gordon Allport, that such an endeavor is the personologist’s *raison d’être*” (McAdams, 1988, p. 1).

In terms of clinical applications, it should be borne in mind that the personological tradition has tended to focus on relatively normal, nonclinical samples and that within this tradition, life histories have been viewed as relatively veridical accounts of what has happened in a person’s life (Runyan, 1982) or as imaginative reconstructions of the past (McAdams, 1993). It is also important to recognize that personality assessment within the personological paradigm occurs on a different level than traditional psychodiagnostic assessment (Alexander, 1990). Whereas personological assessment attempts to understand current personality functioning in terms of an individual’s own life history, traditional psychodiagnostic assessment is concerned with determining the individual’s membership in a group or class of individuals that differs from other groups in terms of psychopathology.

A major difference between the two approaches is that the results of the former [personological] can easily be directed toward answering the questions intended by the latter [psychodiagnostic]. The reverse is, unfortunately, a low probability event in anything other than a global sense. To designate someone as obsessive or hysteric, or depressive, or schizophrenic will place the focus on particular salient aspects of functioning but say little about the dynamics in that individual leading to that particular form of functioning. (Alexander, 1990, p. 7)

4.12.4 MULTIVARIATE TRADITION

4.12.4.1 Historical Background

Sir Francis Galton’s (1888) method for analyzing the “co-relations” between twin pairs was refined by Pearson (1896) and utilized effectively by Spearman (1904) in identifying a “general factor” of intelligence (*g*) that appeared to “underlie” various tests of mental abilities. Factors are statistical abstractions that summarize the relations among test scores in terms of underlying or “latent” variables which may be thought of as operating at several levels. For example, the intelligence quotient (IQ) could be considered to be a general higher-order factor that summarizes relations between verbal and performance major group factors, which in turn, summarize relations among multiple minor group factors (Vernon, 1950).

Raymond B. Cattell, the founding father of the multivariate tradition of personality assessment, was a student of Spearman and was among the first to apply factor-analytic methodology to the study of temperament (e.g., Cattell, 1933). Using cluster-analytic procedures, Cattell (1943) reduced Allport and Odbert’s (1936) exhaustive compilation of trait-descriptive terms from Webster’s unabridged dictionary to 35 “surface traits” (see John, 1990). He administered the 35 surface-trait variables (e.g., talkative vs. silent) to respondents in a series of peer-rating studies and eventually concluded that no less than 16 “source traits” could account for the interrelations among the 35 surface clusters (Cattell, 1949).

The Guilfords (1936, 1939) were also pioneers in the application of factor analysis to personality data, and they assembled evidence for the four interpretable factors of “shyness,” “rhythymia,” “depression,” and “liking for thinking” which would now be called “extraversion,” “conscientiousness,” “neuroticism,” and “intellect,” and which (together with “agreeableness”) constitute the currently important five-factor model (FFM) of personality (Digman, 1996). A decade later, Eysenck (1947) factored the intercorrelations among presenting symptoms of psychiatric patients and found two factors (extraversion and neuroticism) that had both theoretical (Jung, 1971) and empirical (MacKinnon, 1944) precedents (see Eysenck, 1990, pp. 95–98). By the 1950s, the systems of Eysenck (1953), Cattell (1957), and Guilford (1959) had become extensive alternative theories of personality structure.

The history of the multivariate tradition is a curious one that has recently been reconstructed

by Digman (1996). Briefly, Digman argues that as early as the 1930s, there was evidence for the generality and replicability of a FFM of personality that consisted of surgency/extraversion (E), agreeableness (A), conscientiousness (C), neuroticism (N), and intellect/openness (O). Over a period of almost 50 years, a series of cumulative studies may be traced (e.g., Fiske, 1949; Norman, 1963; Tupes & Christal, 1961) that led to an emerging consensus on the utility and generalizability of this five-factor representation (e.g., Digman, 1979; Goldberg, 1977; Wiggins, 1973b). By the 1980s, there were three distinctive interpretations of the nature of these five factors (Costa & McCrae, 1985; Goldberg, 1981; Hogan, 1986), that were followed in the 1990s by still more interpretive perspectives (e.g., Buss, 1996; Wiggins & Trapnell, 1996). It should be noted, however, that although there are a number of workers within the multivariate paradigm who agree on the number, if not the nature, of the factors in the FFM, there are many who strongly contest both the number of factors and their significance (e.g., Block, 1995; Eysenck, 1992; Hough, 1992; Waller & Ben-Porath, 1987).

4.12.4.2 Conceptual Framework

Cattell's (1957) master plan for constructing a theory of personality presupposed a rigorous and representative sampling of the language personality sphere that, in effect, defined the universe of content of behaviors to be explained (Wiggins, 1984). To make certain that this universe of content was properly defined, Norman (1967) replicated the earlier lexical research of Allport and Odbert (1936), using a more recent unabridged dictionary, and subsequently employing more rigorous clustering procedures than those used by Cattell (see John, 1990). Norman's taxonomy of personality attributes was subsequently employed by Goldberg (1981, 1990, 1993) in a program of methodologically elegant research on the natural language of personality that firmly established what he termed the "Big Five" dimensions identified by earlier investigators.

Goldberg's taxonomic research program revitalized interest in the five-factor model and within a relatively brief period of time a variety of new instruments were developed (e.g., Botwin & Buss, 1989; Costa & McCrae, 1985; Goldberg, 1992; Hogan, 1986; Trapnell & Wiggins, 1990) that measured these five dimensions within an equal variety of conceptual frameworks, for example, evolutionary theory (Buss, 1996), trait theory (McCrae & Costa, 1996), lexical theory (Saucier & Goldberg, 1996), socioanalytic theory (Hogan, 1996), and inter-

personal theory (Wiggins & Trapnell, 1996). Moreover, a series of investigations by Costa and McCrae revealed that instruments employed in the other four traditions of personality assessment—psychodynamic (Myers & McCauley, 1985), personological (Jackson, 1984), empirical (Hathaway & McKinley, 1983), and interpersonal (Wiggins, Trapnell, & Phillips, 1988)—showed meaningful convergences with some or all of the dimensions of the FFM (Wiggins & Trapnell, 1997).

4.12.4.3 Assessment Instruments

The NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992b) is the instrument of choice for measuring the dimensions of the FFM of personality. This test provides global measures (domain scores) for each of the five factors and six more specific measures (facet scores) within each of the five domains: (i) neuroticism (anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability); (ii) extraversion (warmth, gregariousness, assertiveness, activity, excitement-seeking, positive emotions); (iii) openness (fantasy, aesthetics, feelings, actions, ideas, values); (iv) agreeableness (trust, straightforwardness, altruism, compliance, modesty, tender-mindedness); and (v) conscientiousness (competence, order, dutifulness, achievement striving, self-discipline). The "astonishingly fruitful research collaboration" (Block, 1995) of Costa and McCrae has produced scores of publications that document the substantive, structural, and empirical validity of this instrument.

4.12.4.4 Interpretive Principles

There are notable differences in interpretive principles suggested by proponents of different multivariate instruments and these would appear to reflect different claims as to the extensiveness and importance of the "universe of content" measured by each instrument, and the theoretical basis on which that content is interpreted for each instrument. With respect to the former, the differences among alternative theories of personality structure within the multivariate tradition have often reflected preferences for different methods of factoring correlations among the phenotypic variables of personality study. Thus, Cattell, who argued for 16 primary source traits, as well as the originator of the multiple-factor method himself (Thurstone, 1934), have both been accused of "overfactoring" (Digman, 1996); while Eysenck (1992), who disputes anything beyond three factors, has been accused of "underfactoring" (Costa & McCrae, 1992a). That almost half a

century of essentially methodological disagreements may be approaching resolution can be seen in the emerging consensus that the FFM of personality is an appropriate "working model" of the universe of content of personality structure that may (and should) be expanded, as necessary, whenever the incremental validity of additional dimensions are demonstrated (Wiggins, 1996).

The theoretical bases on which a five-factor profile may be interpreted are of greater interest to the practitioner. And, as is true of a number of instruments that originated in other personality assessment traditions, it is possible to interpret the FFM from a variety of theoretical perspectives. The NEO PI-R (Costa & McCrae, 1992b) provides explicit principles for interpreting normal and abnormal behavior from the standpoint of modern trait theory (McCrae & Costa, 1996). The Hogan Personality Inventory (Hogan & Hogan, 1992) provides explicit interpretive principles for personnel selection in the workplace from the standpoint of socio-analytic theory (Hogan, 1996). A new edition of the venerable Sixteen Personality Factor Questionnaire (Cattell, Cattell, & Cattell, 1993) provides interpretive principles for five "global factor scales" (Cattell, 1994) from the standpoint of a long-standing "mainstream" personality theory (Wiggins, 1984); and the possibilities for interpreting other well-established instruments, such as the Personality Research Form (Jackson, 1984), from an FFM perspective, are under active investigation (e.g., Jackson, Paunonen, Fraboni, & Goffin, 1996).

4.12.4.5 Applications and Current Status

Historically, there has been little consensus regarding the relation between the normal dimensions of personality, measured by inventories developed within the multivariate tradition, and the dimensions and/or categories of psychopathology. Recently, however, there has been an increased conceptual and empirical focus on this issue, stimulated in part by the definition of personality disorders in terms of personality traits by the American Psychiatric Association (APA; 1980) and by the availability of new instruments and techniques for addressing this problem (Strack & Lorr, 1994).

It would appear that the global dispositions measured by the NEO PI-R provide a comprehensive assessment of emotional (N), interpersonal (E & A), motivational (C), and cognitive (O) styles that are not measured by traditional clinical instruments. As Costa (1991b) noted, "This portrait can be used in psychotherapy: to formulate a tentative diagnosis; to aid the

clinician's development of empathy; to help select appropriate treatments; to identify the client's strengths; and to anticipate the outcome, duration, and course of therapy" (p. 395). A detailed and perceptive account of how this might be done has been provided by Miller (1991). The notable longitudinal stability of NEO PI-R dimensions in the adult personality (McCrae & Costa, 1990) has led to a revised view of the therapeutic enterprise:

we expect that clients will bring these dispositions with them to psychotherapy, and that the therapist should (a) take them into account when trying to understand the individual and his or her problems, and (b) tailor the therapy to fit the needs and styles of the client. (McCrae, 1991, p. 406)

The most extensive psychodiagnostic application of the FFM in general, and of the NEO PI-R in particular, has been to the diagnosis and understanding of personality disorders (Costa & Widiger, 1994). The definition of these disorders in terms of dysfunctional personality traits, by a committee of the APA (1980), stimulated an unprecedented collaborative effort among psychiatrists, psychologists, and psychometricians to find common grounds for characterizing these disorders. It was found, for example, that both the APA diagnostic criteria and the clinical literature on personality disorders were compatible with descriptions of each disorder generated from the facet and domain scales of the NEO PI-R (see Widiger, Trull, Clarkin, Sanderson, & Costa, 1994, Table 1, p. 42).

Subsequent advances in relating the FFM to the personality disorders are too numerous to summarize, but may be found in current journals, particularly the *Journal of Personality Disorders*. The current literature may be viewed as reflecting yet another milestone in the long-standing relation between the FFM and the multivariate paradigm, which Digman (1994) characterized as follows:

Now, after many years of lying on the closet shelf of personality theory, the model has been dusted off, "as good as new," and appears to be for many researchers ... a very meaningful theoretical structure for organizing the myriad specifics implied by the term *personality*. (p. 13)

4.12.5 EMPIRICAL TRADITION

4.12.5.1 Historical Background

The MMPI (Hathaway & McKinley, 1943) is, and has been for many years, the most widely used inventory of personality and psychopathology. To understand the 50-year history

of this instrument is to understand the history of objective personality assessment for the same time period. As Craik (1986) put it, in his historical survey of personality research methods, “the MMPI came to serve as the centerpiece of this period’s [post-WWII] predominant or mainstream agenda” (p. 21). Within that agenda, one can discern the operation of two separate dialectical processes over time: ongoing disputes between the developers of the MMPI and their critics and bipolar shifts in conceptualization that have occurred within the evolving conceptual and interpretive frameworks of the developers of the MMPI. The ongoing disputes reflect the fact that, because of its prominence, the MMPI came to be associated with such contentious issues as clinical versus statistical prediction (Meehl, 1954) and response styles (Wiggins, 1962). The shifts in conceptualization reflect the virtuosity and intellectual flexibility of the developers of the MMPI who were able to shift from typological categories to continuous trait dimensions; from discriminant validity of differential diagnoses to the construct validity of scales and profiles; and from denigration of self-reports to the canonization of item content (Ben-Porath, 1994).

The dialectical processes just described were enacted by a remarkably diverse group of scientists/practitioners whose program of clinical research had its roots in the unique combination of behavioral, biological and psychometric thinking that prevailed at the University of Minnesota in the 1930s and 1940s (Meehl, 1989). Stark R. Hathaway was an originator of the MMPI, a mentor for successive generations of Minnesota students and a merciless critic of his own instrument (e.g., Hathaway, 1972). Paul E. Meehl—learning theorist, philosopher of science, psychoanalyst, and taxometrician—is a legend in his own time, the principal theorist of the MMPI, and arguably the major figure in the field of personality assessment since the late 1950s. W. Grant Dahlstrom is the Talmudic scholar of the MMPI, whose meticulous organizations of that vast literature have informed and influenced generations of clinicians and researchers (e.g., Dahlstrom & Dahlstrom, 1980; Dahlstrom, Welsh, & Dahlstrom, 1972–1975; Welsh & Dahlstrom, 1956). James N. Butcher is a distinguished editor, scholar, and researcher who is principal author of the revised MMPI (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), the new MMPI content scales (Butcher, Graham, Williams, & Ben-Porath, 1990), the adolescent version of the MMPI (MMPI-A; Butcher et al., 1992), and the MMPI automated clinical report (Butcher, 1993).

4.12.5.2 Conceptual Framework

The rationales underlying four distinguishable strategies of objective (as opposed to projective) test construction are based on differing views regarding the meaning of subjects’ responses to objective test items (Wiggins, 1973b).

(i) *The rational (correspondence) strategy.* This assumes a one-to-one correspondence between a subject’s “report” (“I am anxious”) and a palpable internal state of the subject (the experience of anxiety) (Buchwald, 1961). This assumption was the cornerstone of introspectionism in early experimental psychology and it was the principal rationale underlying the earliest personality inventories (e.g., Woodworth, 1917).

(ii) *The empirical (instrumental) strategy.* This views a subject’s “utterance” as “an intrinsically interesting and significant bit of verbal behavior, the non-test correlates of which must be discovered by empirical means” (Meehl, 1945, p. 297, emphasis added). If it were established empirically that a false response to the item “I am anxious” was more characteristic of hospitalized hysterics than of a normal control group, that particular bit of verbal behavior would become meaningful, and would illustrate the instrumental value of verbal behavior as a tool in psychodiagnostics. This philosophy was evident in the early behavioristic critique of introspectionism in experimental psychology and in later critiques of existing rational personality inventories of the 1930s (Humm & Wadsworth, 1935; Landis & Katz, 1934; Landis, Zubin, & Katz, 1935). The earliest example of a major inventory constructed under an empirical strategy was the Vocational Interest Blank (Strong, 1927), which served as the prototype for the construction of the MMPI.

(iii) *The constructive (substantive) strategy.* This views a subject’s response as a manifestation of an underlying personality construct that is embedded in an interlocking system of laws that relate constructs to one another and to observable properties of the environment (Cronbach & Meehl, 1955). The meaning of a construct is given by the empirical laws into which it enters and the significance of a construct is given by the number of such lawful relations discovered. The construct validity of a personality scale is estimated from “the proportion of test score variance that is attributable to the construct variable” (p. 289).

(iv) *The interpersonal (self-presentational) strategy.* This views a subject’s response as an interpersonal communication between the subject and the tester (or the institution that the tester represents) (Carson, 1969b; Leary, 1957). The meaningfulness of such a communication is

dependent upon both the subject's and the tester's views of the meaning of test responses in a particular testing situation (e.g., hospital admission, vocational guidance). Thus, the subject may mean to communicate a complaint ("I am anxious") and the tester (or the scoring system employed by the institution) may encode this as "neuroticism." The subject may view the testing situation as an opportunity for self-disclosure (Jourard, 1964) and the examiner may view it as an opportunity for impression management (Goffman, 1959). Elsewhere it has been argued that the interpersonal view of test responses has much to commend it, not the least of which is the likelihood that this is the frame of reference which subjects themselves adopt (Wiggins, 1966).

4.12.5.3 Assessment Instruments

Construction of the MMPI began with the generation of a large pool of items that were considered representative of "behaviors of significance to the psychiatrist" (Hathaway & McKinley, 1940). These items were presented in true-false format to several groups of diagnosed psychiatric inpatients (e.g., depressives) and to a large group of normal control subjects. The proportion of patients in a given diagnostic group who responded "true" ("endorsement") to each item (e.g., 80%) was compared with the proportion of normal subjects who had responded "true" to that item (e.g., 10%), in order to identify discriminating items that might be keyed (in this instance, "true") on a clinical scale (e.g., depression scale; Hathaway & McKinley, 1942). Innovative procedures were then applied in the development of three validity scales that served to identify sources of invalidity or "faking" in test responding: a "lie scale" (*L*) with items of high desirability and low endorsement, an "infrequency scale" (*F*) with items of low desirability and low endorsement, and a subtle "correction scale" (*K*) that measures defensiveness when high and self-criticalness when low. The resultant three validity scales, eight clinical scales (hypochondriasis, depression, hysteria, psychopathic deviate, paranoia, psychasthenia, schizophrenia, hypomania), a masculinity/femininity scale, and a social introversion scale constitute the MMPI profile that was normed and standardized on the original sample of normal control subjects.

4.12.5.4 Interpretive Principles

4.12.5.4.1 Empirical strategy

The original purpose of the MMPI was to aid psychiatrists in their assignment of patients to

the traditional psychiatric diagnostic categories of that time. The conservative rationale of the empirical strategy permitted such statements as "This subject's pattern of verbal behavior resembles more closely that of a group of hospitalized schizophrenics than it does that of a normal control group." This fact, together with evidence from other sources, was suggestive of the working diagnosis of "schizophrenia." Unfortunately, the individual clinical scales were not successful in subsequent attempts to discriminate diagnostic groups from normals (Hathaway, 1960), due most likely to lack of power in the original statistical comparisons and to unreliability of both predictor and criterion measures.

4.12.5.4.2 Constructive strategy

Ben-Porath (1994) provides an excellent historical summary of the "reinvention of the MMPI" following the failure of the empirical strategy and of its subsequent "evolution into an omnibus measure of personality" within a construct-oriented perspective. The highlights of this shift over a 50-year period of unprecedented research and clinical work include emphases upon: (i) MMPI profile configurations, rather than on single scales (e.g., Gough 1946); (ii) normal personality correlates of profile types (e.g., Black, 1956); (iii) additional empirically and substantively derived scales that broadened the nomological network of MMPI investigation (e.g., Morey, Waugh, & Blashfield, 1985), and (iv) actuarial prediction systems based on configural profile types (e.g., Marks & Seeman, 1963).

4.12.5.4.3 Interpersonal strategy

Loevinger's (1957) formulation of the substantive component of construct validity provided an even greater contrast to the original empirical perspective of the MMPI. Nevertheless, workers within the MMPI tradition for the most part continued to pursue the other components of construct validity just described. Interest in the "content" of patients' communication of complaints via the MMPI remained minimal until the appearance of Wiggins' (1966) MMPI Content Scales. In terms of Loevinger's distinctions, these content scales appeared to provide: (i) representative coverage of the universe of content of the MMPI item pool (see Johnson, Butcher, Null, & Johnson, 1984), (ii) psychometrically sound measures of 13 dimensions of self-report that were interpretable with reference to previous studies of the factorial structure of the MMPI (Welsh, 1956), and (iii) empirical evidence of convergent and

discriminant validity with reference to psychiatric diagnostic categories (Wiggins, 1966; Payne & Wiggins, 1972). Perhaps of greater interest to the MMPI community was the potential usefulness of these content scales in individual psychodiagnostic appraisal (e.g., Nichols, 1987). Variants of the original content scales are now an official supplement to the clinical scales of MMPI-2 (Butcher et al., 1990) and their incremental contribution to the differential diagnosis of psychopathology appears highly promising (e.g., Ben-Porath, Butcher, & Graham, 1991).

4.12.5.5 Applications and Current Status

The MMPI developed in the 1940s was, by contemporary standards, based on inadequate and outdated norms, replete with offensive and anachronistic items, and patently unsuited to the diagnostic task for which it was devised. The clinical scales were derived by less than optimal contrasted group procedures (Jackson, 1971) that, from the beginning, appeared unreplicable (Benton, 1949). In the absence of a coherent theoretical rationale, the profiles, code-types, and actuarial systems based on these scales have an ambiguous conceptual status (Helmes & Reddon, 1993). Pressures to restandardize, revise, or replace this relic have existed almost from its inception, culminating in a historically important summit meeting (Butcher, 1972) in which the third option was seriously considered (Wiggins, 1973a). Once again, however, the resilient and perdurable Minneapolis Group (Butcher et al., 1989) rose to the challenges from within and without to produce what was optimistically dubbed "MMPI-2" (leaving open the possibility of MMPI-3 in the year 2035).

Although the original MMPI is a disgruntled psychometric critic's dream come true, it is difficult to ignore the past and present status of the MMPI as our most widely used objective personality test. Given what appeared to be the lesser of two evils, the personnel of the MMPI Restandardization Project opted to emphasize continuity with the vast interpretive and empirical database accumulated for the original clinical and validity scales while simultaneously removing inappropriate items, correcting psychometric deficiencies where possible, and generating new content scales. Shorn of its dated and politically incorrect items, the lamblike content of the MMPI-2 is now clothed in the lupine garments of computer technology and state-of-the-art methodology (e.g., uniform *T*-scores, adaptive testing, measurement of variable response inconsistency). MMPI-2 appears to be replacing the original in most clinical

settings (e.g., Webb, Levitt, & Rojdev, 1993), although there continue to be criticisms from both without and within the MMPI community (e.g., Caldwell, 1997).

In 1972, Hathaway concluded: "So, in summary, we are stuck, I think, with the MMPI ... for a dreary while longer, although a prophet may even now be wandering in the psychometric wilderness" (1972 p. 40). There is little doubt that we will be "stuck" with MMPI-2 for many years to come, although the paths of the prophets are now more discernible (e.g., Costa, 1991a; Morey, 1991).

4.12.6 INTERPERSONAL PARADIGM

4.12.6.1 Historical Background

The interpersonal circumplex model had its origins in the writings of Harry Stack Sullivan (1953) who introduced a number of original and radical ideas into the American psychiatry of the 1940s. Although not entirely liberated from Freud's spatial-hydraulic metapsychology, Sullivan was "behavioral" in the sense that his primary emphasis was upon the things that persons do to one another in interpersonal transactions, rather than upon internal processes that persons might "have." His oft-cited definition of personality as "the relatively enduring pattern of recurrent interpersonal situations which characterize a human life" represents an even greater departure from psychoanalytic thought, as well as a unique perspective within personality theory itself.

Sullivan was influenced by Lewin's (1938) field-theoretical conceptualization of contemporaneous, bidirectional influences in "psychological fields," which led him to define the interpersonal situation as the basic unit of observation in psychiatry. This idea does not connote a separate "person" or "persons" who may be considered independently from the complex field (situation) of bidirectional causalities in which they are embedded; nor does it hold out the hope of capturing such complexities as "interaction" terms in analyses of variance.

The Kaiser Foundation Research Project, in which Timothy Leary (1957) was the most pre-eminent investigator, attempted to operationalize Sullivanian concepts in terms of concrete measurement procedures. An ordinary language analysis of clinicians' observations of the things that patients did to each other (and to themselves) in group psychotherapy led to a taxonomy of interpersonal behaviors that appeared to be empirically well-captured by a circular arrangement of interpersonal variables organized around the coordinates of

“dominance” and “affiliation.” LaForge, Leary, Nabors, Coffey, and Freedman (1954) derived the basic trigonometry of what would later be called a “circumplex,” without knowledge of Guttman’s (1954) related work on that topic.

Notable contributions during the 1960s included the conceptualization of both maternal and child behavior within an interpersonal circumplex framework (Schaefer, 1961), a psychometrically sophisticated replication of the circumplex within a clinical population (Lorr & McNair, 1963), and a highly influential integration of the circumplex with the clinical, social, and experimental psychology of that time (Carson, 1969a). Alternative conceptual formulations of the interpersonal model were presented in the 1970s (e.g., Benjamin, 1974; Kiesler, 1979; Wiggins, 1979), and during the 1980s the model was applied to psychotherapy (Anchin & Kiesler, 1982), complementarity (Kiesler, 1983), and interpersonal problems (Horowitz, Rosenberg, Baer, Ureno, & Villaseñor, 1988). Among the contributions of the 1990s have been an updated and psychometrically-sound version of the original interpersonal checklist (Wiggins, 1995), a comprehensive exposition of contemporary interpersonal theory and research (Kiesler, 1996), and a presentation of the impressive variety of contexts in which the interpersonal circumplex model has been applied (Plutchik & Conte, 1997).

4.12.6.2 Conceptual Framework

The radical interpersonalism of Sullivan (1953) may be difficult to comprehend on first exposure because of our ingrained “individualist language” (p. 50) for describing personality as reflecting attributes of a discrete individual who is “separate” from others and from a shared social environment. Theorists within the interpersonal paradigm have attempted to operationalize Sullivan’s conceptualization of personality in ways that avoid this individualistic bias, by defining personality as “nothing more (or less) than the patterned regularities that may be observed in an individual’s relations with other persons, who may be real in the sense of actually being present, real but absent and hence ‘personified’ or ‘illusory’” (Carson, 1969a, p. 26).

On a metatheoretical level, it is helpful to think of the dominance and affiliation axes of the interpersonal circumplex in terms of Bakan’s (1966) concepts of “agency” (mastery, self-assertion, self-expansion) and “communion” (union with a larger social or spiritual

entity), respectively (Wiggins, 1991). These concepts are compatible with Sullivan’s emphasis on the communal context of interpersonal transactions and with his belief in the importance of the social sciences for understanding interpersonal situations. When conjoined with Foa and Foa’s (1974) theory of the exchange of love (communion) and status (agency) in interpersonal transactions, these concepts provide a rich conceptual framework for the measurement and understanding of interpersonal behavior (Wiggins & Trapnell, 1996).

4.12.6.3 Assessment Instruments

4.12.6.3.1 Interpersonal Adjective Scales

The Interpersonal Adjective Scales (IAS; Wiggins, 1995) evolved from a psychological taxonomy of trait-descriptive terms that was developed within the framework of a larger program of collaborative research on language and personality (Goldberg, 1977). Within a representative pool of trait-descriptive adjectives selected from an unabridged dictionary, an “interpersonal domain” was distinguished from other domains, such as characterological, temperamental, and cognitive domains (Wiggins, 1979). Approximately 800 interpersonal adjectives were assigned to the original categories of the interpersonal circumplex (Leary, 1957) on both conceptual and empirical grounds. Using computer-based multivariate procedures it was found that scales based on the original categories failed to meet certain circumplex criteria that could be better met with scales based on the revised categories that now constitute the IAS (Wiggins, 1995). The IAS consists of 64 adjectives that respondents rate for self-descriptive accuracy on an eight-place Likert scale ranging from “extremely inaccurate” to “extremely accurate.” Eight scales, or “octants,” of eight items each, assess the interpersonal dispositions listed in the first column of Table 1. The PA octant, for example, includes items such as “dominant,” “forceful,” and “assertive.”

4.12.6.3.2 Inventory of Interpersonal Problems

Horowitz (1979) transcribed statements of interpersonal problems expressed by psychiatric outpatients in the course of videotaped intake interviews and employed these statements as items in the construction of the Inventory of Interpersonal Problems (IIP), in which respondents are required to indicate the extent to which each of 127 statements is problematic on a five-place Likert scale ranging from “not at

Table 1 Octant scales from three interpersonal assessment instruments.

	<i>IAS Interpersonal Adjective Scales</i>	<i>IAS Inventory of Interpersonal Problems</i>	<i>IAS Impact Message Inventory</i>
PA	Assured–dominant	Domineering	Dominant
BC	Arrogant–calculating	Vindictive	Hostile–dominant
DE	Cold–hearted	Cold	Hostile
FG	Aloof–introverted	Socially avoidant	Hostile–submissive
HI	Unassured–submissive	Nonassertive	Submissive
JK	Unassuming–ingenuous	Exploitable	Friendly–submissive
LM	Warm–agreeable	Overly nurturant	Friendly
NO	Gregarious–extraverted	Intrusive	Friendly–dominant

all” to “extremely” (Horowitz et al., 1988). Subsequently, Alden, Wiggins and Pincus (1990) developed a circumplex version of this inventory (IIP-C) that consists of eight scales, of eight items each, that assess the interpersonal problems listed in the second column of Table 1. The PA octant, for example, includes items such as “I try to control other people too much.”

4.12.6.3 Impact Message Inventory

The Impact Message Inventory (IMI) is a highly original and promising method of assessment that is based on Kiesler’s (1988) theory of interpersonal communication in psychotherapy. The theory postulates that disordered individuals are unaware of the unintended, inappropriate, and ambiguous messages they repetitively “send” to others and that they are therefore confused and distressed by the pattern of negative responses they consistently evoke or “pull” from others. The IMI attempts to identify the location within the interpersonal circumplex of these patterns of negative response evoked in others, as a means of gaining insight into a client’s maladaptive transactional behavior.

Items were generated from a content analysis of free responses to 15 interpersonal vignettes that described characters enacting 15 different interpersonal styles, similar to those found in the first column of Table 1. Respondents were asked to imagine themselves in the company of each of these characters and to record their covert reactions using the stem, “He makes me feel . . .” Content analysis of responses suggested three categories of covert reaction: direct feelings, action tendencies, and perceived evoking messages. The Octant version of the IMI consists of six items for each of the target stimuli listed in the third column of Table 1. For each octant, there are two items for each of the three categories of covert reaction to the target stimulus. Thus, for example, the PA scale includes two items each for direct feelings

(e.g., “bossed around”), action tendencies (e.g., “I want to tell him to give someone else a chance to make a decision”), and perceived evoking messages (e.g., “He thinks he’s always in control of things”). Respondents (e.g., a psychotherapist) are asked to imagine themselves in the company of a particular person (e.g., a psychotherapy patient) and to indicate the extent to which they experience the covert reactions on a four-place scale. Covert reactions to a target person (e.g., feeling “bossed around”) are thus revealing of the personality and the interpersonal impact of that target person (e.g., rigidly dominant, PA).

4.12.6.4 Interpretive Principles

One of the most valuable features of the circumplex model is the opportunity it provides for interpreting interpersonal variables with reference to the geometric principles of a circle (LaForge et al., 1954). The variables of traditional multiscale inventories are typically displayed as “factor lists” (Hogan, 1983) in which the ordering and interrelations among variables lack both conceptual and interpretive significance. In contrast, the trigonometric procedures that can be applied to the variables of an interpersonal circumplex permit descriptions and diagnostic inferences that cannot be generated from traditional scales (Wiggins, Phillips, & Trapnell, 1989).

Some general principles of circumplex interpretation may be illustrated with reference to Figure 1, which presents the IAS profile of a 44-year-old woman who was employed as a senior bank manager. At the bottom of the figure are *T*-scores on the octant variables, expressed with reference to an appropriate normative group. These scores have been plotted on shaded sectors of the circle and they are interpreted as representing eight vectors in two-dimensional space. The mean or average directionality of these eight vectors is of critical diagnostic significance because it determines

the “typological” category to which this woman will be assigned. By trigonometric procedures, this average directionality was determined to be 90°, which falls exactly at the midpoint of the assured-dominant category (PA). Such an angular location is considered to be prototypical of individuals classified as pure “PA types.” Had the location been at 110°, a more hostile manifestation of dominance (BC) would have been suggested; had it been at 65°, a more affiliative expression of dominance (NO) would have been suggested. But the prototypicality of the present classification increases our confidence in asserting that the principal interpersonal style of this woman is one that emphasizes the exercise of power over others in a social context, by such activities as taking charge, making decisions, and winning arguments (Wiggins, 1995, p. 22).

The general shape of the profile in Figure 1 closely resembles the characteristic configuration of IAS profiles that is found in the average profiles of subjects in all IAS typological groups (Wiggins et al., 1989). The characteristic configuration of an IAS profile is one in which the principal elevation occurs on the defining octant (in this case, PA), with secondary and approximately equal elevations occurring on adjacent octants (BC and NO), and diminishing and approximately equal elevations occurring on subsequent pairs of octants (DE and LM), (FG and JK), down to a highly truncated “opposite” (to the principal) octant (HI). Thus, across interpersonal situations, we would expect this woman frequently to behave in a forceful, assertive, dominant, and self-confident manner (PA); to somewhat less frequently behave in aggressive (BC) and gregarious (NO) ways; to seldom, if ever, behave in a submissive (HI) fashion, and so forth.

The vector length of an interpersonal profile is a measure of “deviance” in both a statistical and a psychiatric sense. In the former sense, vector length is the standard deviation of the eight interpersonal variables that indicates the “intensity” with which a pattern of interpersonal behavior is expressed. In the latter sense, high-variance profiles are associated with interpersonal “rigidity” and are found most often in psychiatric groups. The length of the arrow indicating angular location in Figure 1 is the vector length of the profile, which is approximately 1.8 standard deviations above that of the normative group. This would suggest that the pattern of interpersonal behaviors described above would be expressed vigorously, and quite possibly rigidly or intemperately.

Having established that this individual’s profile is representative of those obtained by assured dominant “types,” and that this pattern

of behavior is expressed in a clearly differentiated fashion, interpretation would proceed with reference to the empirical literature of the interpersonal tradition that has examined both the dimension of assured-dominant (PA) and the characteristics of dominant “types” in both normal and psychiatric populations (Kiesler, 1996). This would include the literature of other instruments that have studied this location (PA) in contexts such as interpersonal problems (IIP-C), impact messages (IMI), and other contexts that have been studied by interpersonal researchers.

4.12.6.5 Applications and Current Status

The definition of the personality disorders in terms of personality traits by the APA (1980), and the availability of new instruments and techniques for assessing personality disorder, has resulted in an increased understanding of this diagnostic axis. For example, many of the *Diagnostic and statistical manual of mental disorders (DSM)* personality disorder categories have been shown to be well captured by the two-dimensional structures of the IAS (e.g., Wiggins & Pincus, 1989, 1994) and the IIP-C (e.g., Pincus & Wiggins, 1990; Soldz, Budman, Demby, & Merry, 1993). Similarly, clinicians’ ratings of *DSM* personality disorder categories have been found to be well captured by an interpersonal circumplex model (e.g., Blashfield, Sprock, Pinkston, & Hodgin, 1985; Plutchik & Conte, 1986). And more recently, Benjamin (1995) has provided a detailed and perceptive description of procedures for the diagnosis of personality disorders within the framework of her variant of the interpersonal circumplex model.

The interpersonal tradition originated in the context of psychotherapy and, perhaps more than any other tradition, has contributed to an understanding of the therapeutic process itself (e.g., Anchin & Kiesler, 1982; Benjamin, 1995; Kiesler, 1988; Safran & Segal, 1990; Sullivan, 1953). The most notable contribution of this tradition to psychotherapy has been the circumplex structural model that provides a framework for representing dyadic interactions in the therapeutic relationship. A number of different assessment devices based on the interpersonal circumplex, and variants of that model (e.g., Benjamin, 1974), have proven useful in studies of psychotherapy process and outcome (Kiesler, 1996).

Gurtman (1996) has emphasized the construct validity of the circumplex version of Horowitz’s Inventory of Interpersonal Problems (IIP-C; Alden et al., 1990) within the psychotherapy context. Henry (1996) has made

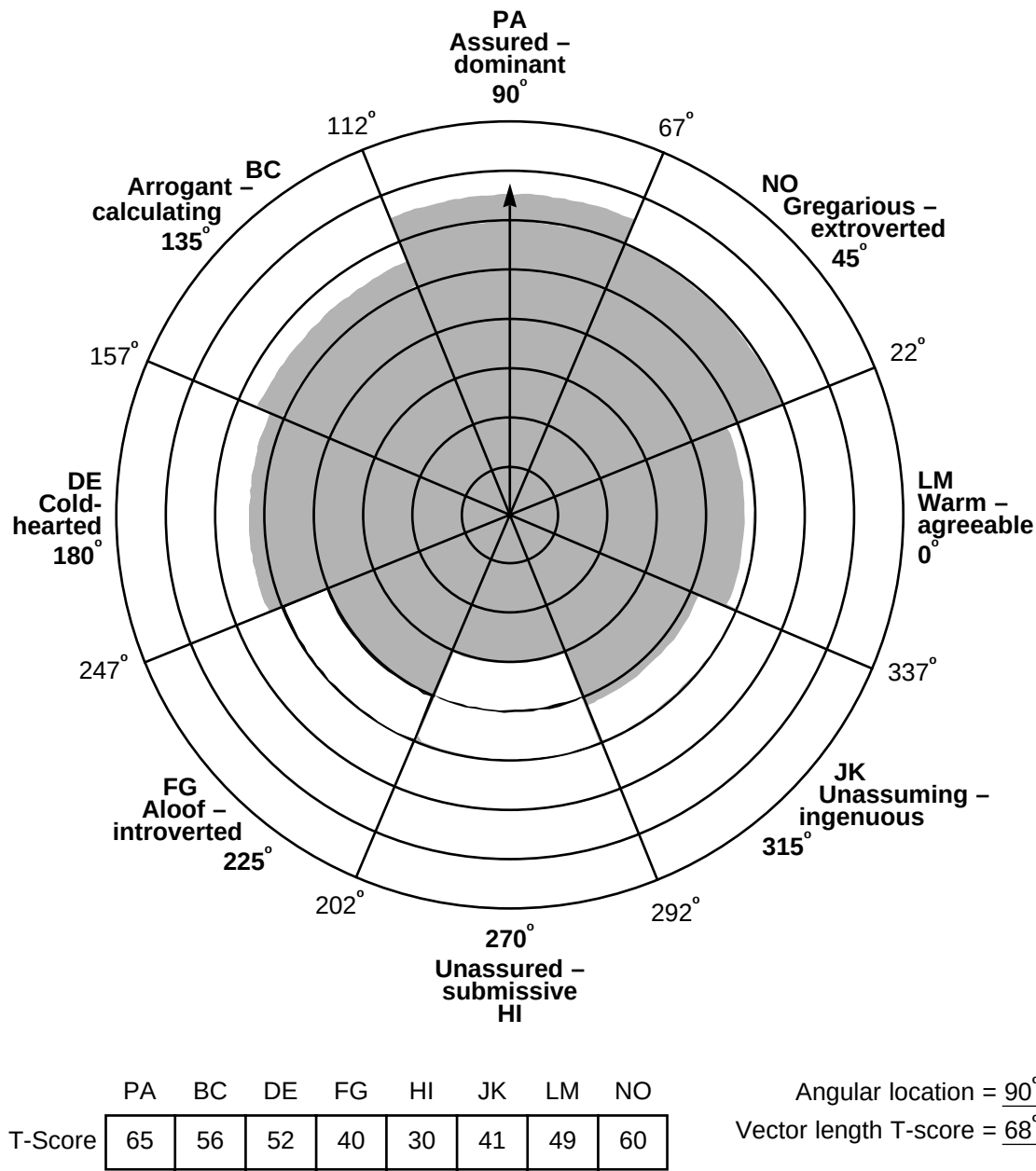


Figure 1 IAS profile of a 44-year-old bank manager.

a similar case for the Structural Analysis of Social Behavior (Benjamin, 1974) as a common metric for programmatic psychotherapy research. Kiesler, Schmidt, and Wagner (1997) have stressed the conceptual advantages of the IMI in analyzing the psychotherapeutic relationship. And Kiesler (1996) has also emphasized the potential of the revised Check List of Psychotherapy Transactions (Kiesler, Goldston, & Schmidt, 1991) for measuring the interpersonal behavior of interactants in the therapy relationship. Taken together, the studies summarized in the aforementioned review articles constitute an impressive empirical

literature that attests to the heuristic potential of the interpersonal circumplex model in psychotherapy research. The future of the interpersonal circumplex tradition for research in both psychodiagnostics and psychotherapy appears to be a bright one (Wiggins & Trobst, 1997).

4.12.7 CONCLUSION

In the introduction to this chapter, we considered the diagnostic work of a hypothetical clinician who was both well trained and

well informed about personality tests, but who was perhaps unaware of the conceptual and interpretive richness, and particularly of the distinctiveness, of the five personality assessment traditions just considered. Assuming that she persevered in reading through this highly condensed overview, there are a few "take home" messages that might now be conveyed to her.

First, in Kelly's (1955) terms, the five traditions arose and were developed within different "foci of convenience" and therefore are often most usefully applied within their specific areas of focus. The psychodynamic tradition of Freud arose from the revolutionary idea that the causes and reasons for behavior are frequently not apparent and that we must look "within" the individual, as it were, to discover the sources (often historical) of contemporary problems of living. The personological tradition arose in response to the question of "How shall a psychological life history be written?" The multivariate tradition arose as an extension of the multifactor assessment of intelligence to the personality domain, and was subsequently concerned with the determination of the basic descriptive dimensions of personality. The empirical tradition arose in the context of descriptive psychiatry, and was originally most concerned with the differential diagnosis of psychopathology. The interpersonal tradition was originally focused on the character and quality of interpersonal relationships and on what persons did to one another within such relationships.

Second, instruments developed or employed within the five traditions have been found to have different "ranges of convenience" (Kelly, 1955), that is, to be generally or specifically applicable to a particular spectrum of assessment problems. Here it must be reiterated that there is not a one-to-one relation between instruments and traditions in this respect. A few examples must suffice: The Rorschach test may be useful in assessing cognitive and perceptual deficits; the WAIS may be useful in detecting covert motivational conflicts; the MMPI item pool has proven a fruitful source for the construction of literally hundreds of scales serving different assessment purposes; the TAT, and its associated need-press framework, has proven useful within many different traditions, as has the NEO PI-R, and the IAS.

Finally, our hypothetical clinician should be aware of the armamentarium of conceptual orientations and instruments that may be brought to bear on a specific referral or assessment task. It is extremely unlikely that there is one test, or one orientation, that will be optimal for all assessment tasks. Practical

limitations, especially managed care in the USA, often preclude the once-favored idea of a "standard battery" of tests. It would also not be fair to expect our hypothetical clinician to master all available orientations and tests. But perhaps she could become familiar with a few more. Attending a workshop or two (outside of her present area of expertise) might enhance the professional life of our hypothetical clinician.

ACKNOWLEDGMENTS

We would like to express our indebtedness to Yossef S. Ben-Porath, Sidney J. Blatt, Dan P. McAdams, David S. Nichols, and William McKinley Runyan for their helpful suggestions concerning earlier drafts of sections within their own areas of expertise. However, we accept full responsibility for any errors of fact or interpretation that remain in the present version.

4.12.8 REFERENCES

- Alden, L. E., Wiggins, J. S., & Pincus, A. L. (1990). Construction of circumplex scales for the Inventory of Interpersonal Problems. *Journal of Personality Assessment*, 53, 521-536.
- Alexander, I. E. (1990). *Personology: Method and content in personality assessment and psychobiography*. Durham, NC: Duke University Press.
- Allison, J., Blatt, S. J., & Zimet, C. N. (1968). *The interpretation of psychological tests*. New York: Harper and Row (Reprinted in 1988 by Taylor & Francis).
- Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt, Rinehart, & Winston.
- Allport, G. W. (1942). *The use of personal documents in psychological science*. New York: Social Science Research Council.
- Allport, G. W. (1965). *Letters from Jenny*. New York: Harcourt, Brace, & World.
- Allport, G. W. (1967). Autobiography. In E. G. Boring & G. Lindzey (Eds.), *A history of psychology in autobiography* (Vol. 5, pp. 1-25). New York: Appleton-Century-Crofts.
- Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycho-lexical study. *Psychological Monographs*, 47 (Whole Issue No. 211).
- American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- Anchin, J. C., & Kiesler, D. J. (Eds.) (1982). *Handbook of interpersonal psychotherapy*. Elmsford, NY: Pergamon.
- Babcock, H. (1933). *A short form of the Babcock examination for the measurement of mental deterioration*. Chicago: Stoelting.
- Bakan, D. (1966). *The duality of human existence: Isolation and communion in Western man*. Boston: Beacon Press.
- Behrends, R. S., & Blatt, S. J. (1985). Internalization and psychological development through the life cycle. *Psychoanalytic Study of the Child*, 40, 11-39.
- Benjamin, L. S. (1974). Structural analysis of social behavior. *Psychological Review*, 81, 392-425.
- Benjamin, L. S. (1995). *Interpersonal diagnosis and treatment of personality disorders* (2nd ed.). New York: Guilford Press.
- Ben-Porath, Y. S. (1994). The MMPI and MMPI-2: Fifty

- years of differentiating normal and abnormal personality. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 361–401). New York: Springer.
- Ben-Porath, Y. S., Butcher, J. N., & Graham, J. R. (1991). Contribution of the MMPI-2 content scales to the differential diagnosis of psychopathology. *Psychological Assessment*, 3, 634–640.
- Benton, A. L. (1949). Review of Minnesota Multiphasic Personality Inventory. In O. K. Buros (Ed.), *The third mental measurements yearbook* (pp. 104–107). Highland Park, NJ: Gryphon Press.
- Black, J. D. (1956). Adjectives associated with various MMPI codes. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine* (pp. 151–172). Minneapolis, MN: University of Minnesota Press.
- Blashfield, R., Sprock, J., Pinkston, K., & Hodgin, J. (1985). Exemplar prototypes of personality disorder diagnoses. *Comprehensive Psychiatry*, 26, 11–21.
- Blatt, S. J., & Allison, J. (1981). The intelligence test in personality assessment. In A. I. Rabin (Ed.), *Assessment with projective techniques* (pp. 137–231). New York: Springer.
- Blatt, S. J., Brenneis, C. B., Schimek, J. G., & Glick, M. (1976). The normal development and psychopathological impairment of the concept of the object on Rorschach. *Journal of Abnormal Psychology*, 85, 364–373.
- Blatt, S. J., & Lerner, H. D. (1983). The psychological assessment of object representation. *Journal of Personality Assessment*, 47, 7–28.
- Block, J. (1995). A contrarian view of the five-factor approach to personality. *Psychological Bulletin*, 117, 187–215.
- Botwin, M. D., & Buss, D. M. (1989). Structure of act-report data: Is the five-factor model of personality recaptured? *Journal of Personality and Social Psychology*, 56, 988–1001.
- Bowlby, J. (1973). *Attachment and loss, Vol. 2: Separation*. New York: Basic Books.
- Buchwald, A. M. (1961). Verbal utterances as data. In H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science* (pp. 461–468). New York: Holt.
- Buss, D. M. (1996). Social adaptation and five major factors of personality. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 180–207). New York: Guilford Press.
- Butcher, J. N. (Ed.) (1972). *Objective personality assessment: Changing perspectives*. New York: Academic Press.
- Butcher, J. N. (1993). *User's guide for the Minnesota Clinical Report*. Minneapolis, MN: National Computer Systems.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
- Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 content scales*. Minneapolis, MN: University of Minnesota Press.
- Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Minnesota Multiphasic Personality Inventory (MMPI-A): Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Caldwell, A. B. (1997). Whither goest our redoubtable mentor, the MMPI/MMPI-2? *Journal of Personality Assessment*, 68, 47–68.
- Carson, R. C. (1969a). *Interaction concepts of personality*. Chicago: Aldine.
- Carson, R. C. (1969b). *Interpretive manual to the MMPI*. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications* (pp. 41–53). New York: McGraw-Hill.
- Cattell, H. E. P. (1994). Development of the 16 PF fifth edition. In S. R. Conn & M. L. Rieke (Eds.), *The 16 PF fifth edition technical manual* (pp. 3–20). Champaign, IL: Institute of Personality and Ability Testing.
- Cattell, R. B. (1933). Temperament tests. II: Tests. *British Journal of Psychology*, 23, 308–329.
- Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, 38, 476–506.
- Cattell, R. B. (1949). *The Sixteen Personality Factor Questionnaire* (1st ed.). Champaign, IL: Institute of Personality and Ability Testing.
- Cattell, R. B. (1957). *Personality and motivation structure and measurement*. Yonkers-on-Hudson, NY: World Book.
- Cattell, R. B., Cattell, A., & Cattell, H. E. P. (1993). *Sixteen Personality Factor Questionnaire* (5th ed.). Champaign, IL: Institute of Personality and Ability Testing.
- Costa, P. T., Jr. (Ed.) (1991a). Clinical use of the five-factor model of personality [Special series]. *Journal of Personality Assessment*, 57(3), 393–464.
- Costa, P. T., Jr. (1991b). Clinical use of the five-factor model: An introduction. *Journal of Personality Assessment*, 57, 393–398.
- Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., Jr., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences*, 13, 653–665.
- Costa, P. T., Jr., & McCrae, R. R. (1992b). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., Jr., & Widiger, T. A. (Eds.) (1994). *Personality disorders and the five-factor model of personality*. Washington, DC: American Psychological Association.
- Craik, K. H. (1986). Personality research methods: An historical perspective. *Journal of Personality*, 54, 18–51.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Dahlstrom, W. G., & Dahlstrom, L. E. (Eds.) (1980). *Basic readings on the MMPI: A new selection on personality measurement*. Minneapolis, MN: University of Minnesota Press.
- Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972–1975). *An MMPI handbook*. (Vols. 1 & 2). Minneapolis, MN: University of Minnesota Press.
- Digman, J. M. (1979, November). *The five major domains of personality variables: Analysis of questionnaire data in light of the five robust factors emerging from studies of rated characteristics*. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Los Angeles, CA.
- Digman, J. M. (1994). Historical antecedents of the five-factor model. In P. T. Costa, Jr. & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 13–18). Washington, DC: American Psychological Association.
- Digman, J. M. (1996). The curious history of the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 1–20). New York: Guilford Press.
- Elms, A. C. (1994). *Uncovering lives: The uneasy alliance of biography and psychology*. New York: Oxford University Press.
- Erikson, E. H. (1950). *Childhood and society*. New York: Norton.

- Erikson, E. H. (1958). *Young man Luther: A study in psychoanalysis and history*. New York: Norton.
- Erikson, E. H. (1969). *Ghandi's truth: On the origins of militant nonviolence*. New York: Norton.
- Eysenck, H. J. (1947). *Dimensions of personality*. London: Routledge & Kegan Paul.
- Eysenck, H. J. (1953). *The structure of human personality*. New York: Wiley.
- Eysenck, H. J. (1990). *Rebel with a cause*. London: Allen.
- Eysenck, H. J. (1992). Four ways five-factors are not basic. *Personality and Individual Differences*, 6, 667–673.
- Fairbairn, W. R. D. (1952). *Psychoanalytic studies of the personality: The object relation theory of personality*. London: Routledge & Kegan Paul.
- Fiske, D. W. (1949). Consistency of the factorial structure of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, 44, 329–344.
- Foa, U. G., & Foa, E. B. (1974). *Societal structures of the mind*. Springfield, IL: Charles C. Thomas.
- Galton, F. (1888). Co-relations and their measurement. *Proceedings of the Royal Society*, 45, 135–145.
- Gill, M. M. (Ed.) (1967). *The collected papers of David Rapaport*. New York: Basic Books.
- Gill, M. M., & Klein, G. S. (1967). The structuring of drive and reality: David Rapaport's contributions to psychoanalysis and psychology. In M. M. Gill (Ed.), *The collected papers of David Rapaport* (pp. 8–34). New York: Basic Books.
- Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY: Doubleday Anchor.
- Goldberg, L. R. (1977, August). *Language and personality: Developing a taxonomy of trait-descriptive terms*. Invited address to the Division of Evaluation and Measurement at the 86th annual meeting of the American Psychological Association, San Francisco.
- Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.
- Goldberg, L. R. (1990). An alternative "Description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59, 1216–1229.
- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4, 26–34.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34.
- Goldstein, K., & Scheerer, M. (1941). Abstract and concrete behavior: An experimental study with special tests. *Psychological Monographs*, 53(2) (Whole Issue No. 239).
- Gough, H. G. (1946). Diagnostic patterns on the MMPI. *Journal of Clinical Psychology*, 2, 23–37.
- Guilford, J. P. (1959). *Personality*. New York: McGraw-Hill.
- Guilford, J. P., & Guilford, R. B. (1936). Personality factors S, E, and M, and their measurement. *Journal of Personality*, 34, 21–36.
- Guilford, J. P., & Guilford, R. B. (1939). Personality factors D, R, T, and A. *Journal of Abnormal and Social Psychology*, 34, 21–36.
- Gurtman, M. B. (1996). Interpersonal problems and the psychotherapy context: The construct validity of the Inventory of Interpersonal Problems. *Psychological Assessment*, 8, 241–255.
- Guttman, L. (1954). A new approach to factor analysis: The radex. In P. R. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). Glencoe, IL: Free Press.
- Hall, C. S., & Lindzey, G. (1978). *Theories of personality* (3rd ed.). New York: Wiley.
- Hanfmann, E., & Kasanin, J. (1937). A method for the study of concept formation. *Journal of Psychology*, 3, 521–540.
- Hartmann, H. (1939). *Ego psychology and the problem of adaptation*. New York: International Universities Press.
- Hathaway, S. R. (1960). Forward. In W. G. Dahlstrom, G. S. Welsh, & L. E. Dahlstrom (Eds.), *An MMPI handbook: A guide to use in practice and research* (pp. vii–xi). Minneapolis, MN: University of Minnesota Press.
- Hathaway, S. R. (1972). Where have we gone wrong? The mystery of the missing progress. In J. N. Butcher (Ed.), *Objective personality assessment: Changing perspectives* (pp. 21–43). New York: Academic Press.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I Construction of the schedule. *Journal of Psychology*, 10, 249–254.
- Hathaway, S. R., & McKinley, J. C. (1942). A multiphasic personality schedule (Minnesota): III. The measurement of symptomatic depression. *Journal of Psychology*, 14, 73–84.
- Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis, MN: University of Minnesota Press.
- Hathaway, S. R., & McKinley, J. C. (1983). *The Minnesota Multiphasic Personality Inventory manual*. New York: Psychological Corporation.
- Helmes, E., & Reddon, J. R. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin*, 113, 453–471.
- Henry, W. P. (1996). The structural analysis of social behavior as a common metric for programmatic psychotherapy research. *Journal of Consulting and Clinical Psychology*, 64, 1263–1275.
- Hogan, R. (1983). A socioanalytic theory of personality. In M. M. Page (Ed.), *1982 Nebraska symposium on motivation: Personality—current theory and research* (pp. 55–89). Lincoln, NE: University of Nebraska Press.
- Hogan, R. (1986). *Hogan Personality Inventory manual*. Minneapolis, MN: National Computer Systems.
- Hogan, R. (1994). Reinventing ourselves (Review of D. P. McAdams, *The stories we live by*). *Contemporary Psychology*, 39, 355–356.
- Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 163–179). New York: Guilford Press.
- Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
- Holt, R. R. (1951). The Thematic Apperception Test. In H. A. Anderson & G. L. Anderson (Eds.), *An introduction to projective techniques* (pp. 181–229). New York: Prentice-Hall.
- Holt, R. R. (Ed.) (1967). Motives and thought: Psychoanalytic essays in honor of David Rapaport. *Psychological Issues*, 5, Monograph 18/19.
- Holt, R. R., and Havel, J. (1960). A method for assessing primary and secondary process in Rorschach responses. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology* (pp. 263–315). New York: Wiley.
- Horowitz, L. M. (1979). On the cognitive structure of interpersonal problems treated in psychotherapy. *Journal of Consulting and Clinical Psychology*, 47, 5–15.
- Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., & Villasenor, V. S. (1988). The Inventory of Interpersonal Problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, 56, 885–892.
- Hough, L. (1992). The "Big Five" personality variables—construct confusion: Description versus prediction. *Human Performance*, 5, 139–155.
- Humm, D. G., & Wadsworth, G. W. (1935). The Humm-

- Wadsworth temperament scale. *Journal of Applied Psychology*, 92, 163–200.
- Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, 78, 229–248.
- Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Fort Huron, MI: Research Psychologists Press.
- Jackson, D. N., Paunonen, S. V., Fraboni, M., & Goffin, R. D. (1996). A five-factor versus six-factor model of personality structure. *Personality and Individual Differences*, 20, 33–45.
- Jacobson, E. (1954). *The self and the object world*. New York: International Universities Press.
- John, O. P. (1990). The “Big Five” factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.
- Johnson, J. H., Butcher, J. N., Null, C., & Johnson, K. N. (1984). Replicated item level factor analysis of the full MMPI. *Journal of Personality and Social Psychology*, 47, 105–114.
- Jourard, S. M. (1964). *The transparent self*. Princeton, NJ: Van Nostrand.
- Jung, C. G. (1971). *Psychological types* (H. G. Baynes, Trans.; revised by R. F. C. Hull). Princeton, NJ: Princeton University Press (Original work published 1923).
- Kelly, G. A. (1955). *The psychology of personal constructs* (2 vols.). New York: Norton.
- Kiesler, D. J. (1979). An interpersonal communication analysis of relationship in psychotherapy. *Psychiatry*, 42, 299–311.
- Kiesler, D. J. (1983). The 1982 Interpersonal Circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90, 185–214.
- Kiesler, D. J. (1988). *Therapeutic metacommunication: Therapist impact disclosure as feedback in psychotherapy*. Palo Alto, CA: Consulting Psychologists Press.
- Kiesler, D. J. (1996). *Contemporary interpersonal theory and research: Personality, psychopathology, and psychotherapy*. New York: Wiley.
- Kiesler, D. J., Goldston, C. S., & Schmidt, J. A. (1991). *Manual for the Check List of Interpersonal Transactions-Revised (CLOIT-R) and the Check List of Psychotherapy transactions-revised (CLOPT-R)*. Richmond, VA: Virginia Commonwealth University.
- Kiesler, D. J., & Schmidt, J. A. (1993). *The Impact Message Inventory: Form IIA Octant Scale version*. Palo Alto, CA: Mind Garden.
- Kiesler, D. J., Schmidt, J. A., & Wagner, C. C. (1997). A circumplex inventory of impact messages: An operational bridge between emotion and interpersonal behavior. In R. Plutchik & H. R. Conte (Eds.), *Circumplex models of personality and emotions* (pp. 221–244). Washington, DC: American Psychological Association.
- Kluckhohn, C., & Murray, H. A. (1953). Personality formation: The determinants. In C. Kluckhohn, H. A. Murray, & D. M. Schneider (Eds.), *Personality in nature, society, and culture* (pp. 53–67). New York: Knopf.
- Kohut, H. (1971). *The analysis of the self: A systematic psychoanalytic approach to the treatment of narcissistic personality disorders*. New York: International Universities Press.
- LaForge, R., Leary, T. F., Naboisek, H., Coffey, H. S., & Freedman, M. B. (1954). The interpersonal dimension of personality: II. An objective study of repression. *Journal of Personality*, 23, 129–153.
- Landis, C., & Katz, S. E. (1934). The validity of certain questions which purport to measure neurotic tendencies. *Journal of Applied Psychology*, 18, 343–356.
- Landis, C., Zubin, J., & Katz, S. E. (1935). Empirical validation of three personality adjustment inventories. *Journal of Educational Psychology*, 26, 321–330.
- Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald.
- Lewin, K. (1938). *The conceptual representation and measurement of psychological forces*. Durham, NC: Duke University Press.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Lorr, M., & McNair, D. M. (1963). An interpersonal behavior circle. *Journal of Abnormal and Social Psychology*, 67, 68–75.
- MacKinnon, D. W. (1944). The structure of personality. In J. McV. Hunt (Ed.), *Personality and the behavior disorders* (Vol. 1, pp. 3–48). New York: Ronald.
- MacKinnon, D. W. (1975). IPAR's contribution to the conceptualization and study of creativity. In I. A. Taylor & J. W. Getzels (Eds.), *Perspectives in creativity* (pp. 60–89). Chicago: Aldine.
- Main, M., Kaplan, N., & Cassidy, J. (1985). Security in infancy, childhood, and adulthood: A move to the level of representation. *Monographs of the Society for Research in Child Development*, 50, 66–104.
- Marks, P. A., & Seeman, W. (1963). *The actuarial description of abnormal personality: An atlas for use with the MMPI*. Baltimore, MD: Williams & Wilkins.
- Mayman, M. (1967). Object representations and object relationships in Rorschach responses. *Journal of Projective Techniques*, 31, 17–25.
- McAdams, D. P. (1988). Biography, narrative, and lives: An introduction. *Journal of Personality*, 56, 1–18.
- McAdams, D. P. (1989). *Intimacy: The need to be close*. New York: Doubleday.
- McAdams, D. P. (1993). *The stories we live by: Personal myths and the making of the self*. New York: William Morrow.
- McAdams, D. P. (1994). *The person: An introduction to personality psychology* (2nd ed.). Fort Worth, TX: Harcourt Brace.
- McAdams, D. P. (1996). Personality, modernity, and the storied self: A contemporary framework for studying persons (Target article). *Psychological Inquiry*, 7, 295–321.
- McClelland, D. C. (1961). *The achieving society*. New York: Van Nostrand.
- McCrae, R. R. (1991). The five-factor model and its assessment in clinical settings. *Journal of Personality Assessment*, 57, 399–414.
- McCrae, R. R., & Costa, P. T., Jr. (1990). *Personality in adulthood: Emerging lives, enduring dispositions*. New York: Guilford Press.
- McCrae, R. R., & Costa, P. T., Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 51–87). New York: Guilford Press.
- Meehl, P. E. (1945). The dynamics of “structured” personality tests. *Journal of Clinical Psychology*, 1, 296–303.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis, MN: University of Minnesota Press.
- Meehl, P. E. (1989). Autobiography. In G. Lindzey (Ed.), *A history of psychology in autobiography* (Vol. 3, pp. 337–389). Stanford, CA: Stanford University Press.
- Miller, T. R. (1991). The psychotherapeutic utility of the five-factor model of personality: A clinician's experience. *Journal of Personality Assessment*, 57, 415–433.
- Millon, T. (1969). *Modern psychopathology*. Philadelphia: Saunders.
- Morey, L. C. (1991). *Personality Assessment Inventory manual*. Odessa, FL: Psychological Assessment Resources.
- Morey, L. C., Waugh, M. H., & Blashfield, R. K. (1985).

- MMPI scales for *DSM-III* personality disorders: Their derivation and correlates. *Journal of Personality Assessment*, 49, 245–251.
- Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: Thematic Apperception Test. *Archives of Neurology and Psychiatry*, 34, 209–306.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Murray, H. A. (1943). *The Thematic Apperception Test: Manual*. Cambridge, MA: Harvard University Press.
- Murray, H. A. (1959). Preparations for the scaffold of a comprehensive system. In S. Koch (Ed.), *Psychology: A study of science* (Vol. 3, pp. 7–54). New York: McGraw-Hill.
- Myers, I. B., & McCauley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.
- Nichols, D. S. (1987). Interpreting the Wiggins MMPI content scales. In K. L. Moreland & J. N. Butcher (Eds.), *Clinical notes on the MMPI* (No. 10, pp. 3–26). Minneapolis: National Computer Systems.
- Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574–583.
- Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics in a university population*. Ann Arbor, MI: University of Michigan, Department of Psychology.
- Office of Strategic Services Assessment Staff (1948). *Assessment of men*. New York: Rinehart.
- Payne, F. D., & Wiggins, J. S. (1972). MMPI profile types and the self-report of psychiatric patients. *Journal of Abnormal Psychology*, 79, 1–8.
- Pearson, K. (1896). Mathematical contributions to the theory of evolution: Regression, heredity, and panmixia. *Philosophical Transactions*, 187a, 253–318.
- Pincus, A. L., & Wiggins, J. S. (1990). Interpersonal problems and conceptions of personality disorders. *Journal of Personality Disorders*, 4, 342–352.
- Plutchik, R., & Conte, H. R. (Eds.) (1997). *Circumplex models of personality and emotions*. Washington, DC: American Psychological Association.
- Plutchik, R., & Conte, H. R. (1986). Quantitative assessment of personality disorders. In R. Michels & J. O. Cavenar, Jr. (Eds.), *Psychiatry* (Vol. 1, pp. 1–13). Philadelphia: Lippincott.
- Prelinger, E., & Zimet, C. N. (1964). *An ego-psychological approach to character assessment*. Glencoe, IL: The Free Press.
- Rapaport, D. (1942). Principles underlying projective techniques. *Character and Personality*, 10, 213–219 (reprinted in Gill, 1967).
- Rapaport, D. (1944–1946). *Manual of diagnostic psychological testing* (2 vols.). New York: Josiah Macy, Jr. Foundation.
- Rapaport, D. (1946). Principles underlying nonprojective tests of personality. *Annals of the New York Academy of Sciences*, 46, 643–652 (reprinted in Gill, 1967).
- Rapaport, D. (1951). The conceptual model of psychoanalysis. *Journal of Personality*, 20, 56–81 (reprinted in Gill, 1967).
- Rapaport, D. (1960). The structure of psychoanalytic theory: A systematizing attempt. *Psychological Issues*, 2, Monograph 6.
- Rapaport, D., Gill, M., & Schafer, R. (1946). *Diagnostic psychological testing* (2 vols.). Chicago: Year Book.
- Rapaport, D., Menninger, K. A., & Schafer, R. (1947). The new role of psychological testing in psychiatry. *American Journal of Psychiatry*, 103, 473–476 (reprinted in Gill, 1967).
- Rorschach, H. (1921). *Psychodiagnostik*. Bern, Switzerland: Bircher. (Trans. Hans Huber Verlag, 1942).
- Runyan, W. McK. (1982). *Life histories and psychobiography: Explorations in theory and method*. New York: Oxford University Press.
- Safran, J. D., & Segal, Z. V. (1990). *Interpersonal process in cognitive therapy*. New York: Basic Books.
- Sarbin, T. R. (Ed.) (1986). *Narrative psychology: The storied nature of human conduct*. New York: Oxford University Press.
- Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 21–50). New York: Guilford Press.
- Schaefer, E. S. (1961). Converging conceptual models for maternal behavior and for child behavior. In J. G. Glidewell (Ed.), *Parental attitudes and child behavior* (pp. 124–146). Springfield, IL: Charles C. Thomas.
- Schafer, R. (1948). *The clinical application of psychological tests*. New York: International Universities Press.
- Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing: Theory and application*. New York: Grune & Stratton.
- Schafer, R. (1967). *Projective testing and psychoanalysis*. New York: International Universities Press.
- Soldz, S., Budman, S., Demby, A., & Merry, J. (1993). Representation of personality disorders in circumplex and five-factor space: Explorations with a clinical sample. *Psychological Assessment*, 5, 53–63.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Stern, G. G., Stein, M. I., & Bloom, B. S. (1956). *Methods in personality assessment*. Glencoe, IL: Free Press.
- Stewart, A. J., Franz, C., & Layton, L. (1988). The changing self: Using personal documents to study lives. *Journal of Personality*, 56, 41–74.
- Strack, S., & Lorr, M. (1994). Introduction. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. xiii–xviii). New York: Springer.
- Strong, E. K., Jr. (1927). *Vocational Interest Blank (Form A)*. Stanford, CA: Stanford University Press.
- Sullivan, H. S. (1953). *The interpersonal theory of psychiatry*. New York: Norton.
- Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 41, 1–32.
- Tomkins, S. S. (1947). *The Thematic Apperception Test*. New York: Grune & Stratton.
- Tomkins, S. S. (1979). Script theory: Differential magnification of affects. In H. E. Howe, Jr. & R. A. Dienstbier (Eds.), *Nebraska symposium on motivation* (Vol. 26, pp. 201–236). Lincoln, NE: University of Nebraska Press.
- Trapnell, P. D., & Wiggins, J. S. (1990). Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. *Journal of Personality and Social Psychology*, 59, 781–790.
- Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings (USAF ASD Tech. Rep. No. 61–97)*. Lackland Air Force Base, TX: US Air Force.
- Vernon, P. E. (1950). *The structure of human abilities*. London: Methuen.
- Waller, N. G., & Ben-Porath, Y. S. (1987). Is it time for clinical psychology to embrace the five-factor model of personality? *American Psychologist*, 42, 887–889.
- Webb, J. T., Levitt, E. E., & Rojdev, R. (1993, March). *After three years: A comparison of the clinical use of the MMPI and MMPI-2*. Paper presented at the 53rd Annual Meeting of the Society for Personality Assessment, San Francisco, CA.
- Wechsler, D. (1941). *The measurement of adult intelligence*. Baltimore: Williams & Wilkins.

- Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams & Wilkins.
- Welsh, G. S. (1956). Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine* (pp. 264–281). Minneapolis, MN: University of Minnesota Press.
- Welsh, G. S., & Dahlstrom, W. G. (1956). *Basic readings on the MMPI in psychology and medicine*. Minneapolis, MN: University of Minnesota Press.
- Westen, D. (1990). Psychoanalytic approaches to personality. In L. Pervin (Ed.), *Handbook of personality theory and research* (pp. 21–65). New York: Guilford Press.
- Westen, D. (1991). Clinical assessment of object relations using the TAT. *Journal of Personality Assessment*, 56, 56–74.
- White, R. W. (1952). *Lives in progress* (1st ed.). New York: Holt, Rinehart, & Winston.
- White, R. W. (1966). *Lives in progress* (2nd ed.). New York: Holt, Rinehart, & Winston.
- White, R. W. (1975). *Lives in progress* (3rd ed.). New York: Holt, Rinehart, & Winston.
- White, R. W. (1981). Exploring personality the long way: The study of lives. In A. I. Rabin, J. Aronoff, A. M. Barclay, & R. A. Zucker (Eds.), *Further explorations in personality* (pp. 3–19). New York: Wiley.
- Widiger, T. A., Trull, T. J., Clarkin, J. F., Sanderson, C., & Costa, P. T., Jr. (1994). A description of the *DSM-III-R* and *DSM-IV* personality disorders with the five-factor model of personality. In P. T. Costa, Jr. & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 41–71). Washington, DC: American Psychological Association.
- Wiggins, J. S. (1962). Strategic, method, and stylistic variance in the MMPI. *Psychological Bulletin*, 59, 224–242.
- Wiggins, J. S. (1966). Substantive dimensions of self-report in the MMPI item pool. *Psychological Monographs*, 80, (22, Whole No. 630).
- Wiggins, J. S. (1973a). Despair and optimism in Minneapolis (A review of J. N. Butcher (Ed.), *Objective personality assessment: Changing perspectives*). *Contemporary Psychology*, 18, 605–606.
- Wiggins, J. S. (1973b). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.
- Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, 37, 395–412.
- Wiggins, J. S. (1984). Cattell's system from the perspective of mainstream personality theory. *Multivariate Behavioral Research*, 19, 176–190.
- Wiggins, J. S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In W. Grove & D. Cicchetti (Eds.), *Thinking clearly about psychology: Essays in honor of Paul E. Meehl* (Vol. 2, pp. 89–113). Minneapolis, MN: University of Minnesota Press.
- Wiggins, J. S. (1995). *Interpersonal Adjective Scales: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Wiggins, J. S. (1996). Preface. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. vii–xi). New York: Guilford.
- Wiggins, J. S. (in press). *Paradigms of personality assessment*. New York: Guilford Press.
- Wiggins, J. S., Phillips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: Evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology*, 56, 296–305.
- Wiggins, J. S., & Pincus, A. L. (1989). Conceptions of personality disorders and dimensions of personality. *Psychological Assessment*, 1, 303–316.
- Wiggins, J. S., & Pincus, A. L. (1992). Personality: Structure and assessment. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 43, pp. 473–504). Palo Alto, CA: Annual Reviews.
- Wiggins, J. S., & Pincus, A. L. (1994). Personality structure and the structure of personality disorders. In P. T. Costa, Jr. & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 73–93). Washington, DC: American Psychological Association.
- Wiggins, J. S., & Trapnell, P. D. (1996). A dyadic-interactional perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 88–162). New York: Guilford.
- Wiggins, J. S., & Trapnell, P. D. (1997). Personality structure: The return of the Big Five. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 737–764). San Diego, CA: Academic Press.
- Wiggins, J. S., & Trapnell, P. D., & Phillips, N. (1988). Psychometric and geometric characteristics of the revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research*, 23, 517–530.
- Wiggins, J. S., & Trobst, K. K. (1997). Prospects for the assessment of normal and abnormal interpersonal behavior. *Journal of Personality Assessment*, 68, 110–126.
- Winter, D. G. (1973). *The power motive*. New York: Free Press.
- Woodworth, R. S. (1917). *Personal data sheet*. Chicago: Stoelting.

4.13

Observations of Parents, Teachers, and Children: Contributions to the Objective Multidimensional Assessment of Youth

DAVID LACHAR

University of Texas-Houston Medical School, Houston, TX, USA

4.13.1 ISSUES IN OBJECTIVE MULTIDIMENSIONAL MULTISOURCE ASSESSMENT	372
4.13.1.1 Overview	372
4.13.1.2 Multidimensional Measurement	372
4.13.1.3 Multisource Assessment	373
4.13.1.4 Psychometric and Technical Considerations	374
4.13.1.4.1 Scale stability and homogeneity	374
4.13.1.4.2 Scale construction methods and scale character	375
4.13.1.4.3 Evidence of scale and questionnaire validity	376
4.13.2 MULTIDIMENSIONAL MULTISOURCE RATING SCALES	377
4.13.2.1 Overview	377
4.13.2.2 <i>Conners' Rating Scales-Revised</i>	378
4.13.2.2.1 <i>Conners' Parent Rating Scale-Revised</i>	378
4.13.2.2.2 <i>Conners' Teacher Rating Scale-Revised</i>	379
4.13.2.2.3 <i>Conners-Wells' Adolescent Self-Report Scale</i>	379
4.13.2.2.4 <i>Commentary</i>	380
4.13.2.3 <i>Child Behavior Checklist/4-18 and 1991 Profile (CBCL); Teacher's Report Form and 1991 Profile (TRF); Youth Self-report and 1991 Profile (YSR)</i>	380
4.13.2.3.1 <i>Child Behavior Checklist/4-18 (CBCL) and 1991 Profile</i>	382
4.13.2.3.2 <i>Teacher's Report Form (TRF) and 1991 Profile</i>	383
4.13.2.3.3 <i>Youth Self-report (YSR) and 1991 Profile</i>	383
4.13.2.3.4 <i>Commentary</i>	384
4.13.2.4 <i>Behavior Assessment System for Children</i>	385
4.13.2.4.1 <i>BASC Parent Rating Scales</i>	386
4.13.2.4.2 <i>BASC Teacher Rating Scales</i>	386
4.13.2.4.3 <i>BASC Self-Report of Personality (SRP)</i>	386
4.13.2.4.4 <i>BASC Monitor for ADHD</i>	387
4.13.2.4.5 <i>Commentary</i>	387
4.13.2.5 <i>Personality Inventory for Children, 2nd Edition (PIC-2); Personality Inventory for Youth (PIY); Student Behavior Survey (SBS)</i>	388
4.13.2.5.1 <i>Personality Inventory for Children, 2nd Edition (PIC-2); Personality Inventory for Youth (PIY)</i>	389
4.13.2.5.2 <i>Student Behavior Survey</i>	397
4.13.2.5.3 <i>Commentary</i>	398

4.13.1 ISSUES IN OBJECTIVE MULTIDIMENSIONAL MULTISOURCE ASSESSMENT

4.13.1.1 Overview

The use of objective questionnaires in the clinical assessment of child and adolescent adjustment continues to grow in popularity. The five years that followed my last review of objective assessment of children and adolescents (Lachar, 1993), have brought additional psychometric advances in instrument development and application. The importance of objective measures in current practice has also been demonstrated by the relative coverage provided in monographs and textbooks available for graduate education and continuing education (Kamphaus & Frick (1996) is an excellent example that supports this observation).

It is also important to consider the effect recent changes in the provision of health care may have upon the use of psychological testing in the diagnostic process. Managed care may influence the use of objective assessment in two quite different ways. First, in the pursuit of cost containment, the use of time-intensive individually administered psychological tests is less likely to be authorized in the routine evaluation of youth, quite independently of the documented value of such procedures. Relatively inexpensive diagnostic questionnaires, in contrast, are more likely to be accepted as a routine component of initial evaluations because they improve upon the efficiency of the intake data collection process. Second, these families of multidimensional measures can contribute substantial value to the program evaluation aspects of managed care. Objective questionnaires have the potential to independently document both the need for treatment as well as to quantify treatment efficacy through repeated application.

This chapter will expand upon many of the themes presented in Lachar (1993) and will describe in some detail four “families” of assessment instruments that meet current standards for intake assessment and treatment evaluation. These materials: (i) assess multiple dimensions of problem behavior, (ii) collect observations from parents, teachers, and youth, and (iii) provide standard scores based upon contemporary national samples. In 1993 only one set of measures met these criteria (Achenbach, 1992); at the conclusion of 1997 four sets of such scales were either commercially available or in press.

4.13.1.2 Multidimensional Measurement

Multidimensional instruments require standardized responses to similarly formatted descriptive statements which generate estimates of a variety of dimensions of child behavior or psychopathology. These estimates are usually visually displayed in the form of a profile of standard scores. These standard scores represent a common metric that has been derived from one normative reference sample or overlapping or similar samples, and are therefore directly comparable (vs. the case of scores from several single-dimension measures that have been standardized on normative samples collected in very different places and times, as well as under different conditions).

Advances in psychopathology and measurement technology demonstrate the clinical advantage of multidimensional measurement. It is the case, rather than the exception, in clinical assessment that several scaled dimensions of adjustment, such as depression and anxiety or inattention and noncompliance, are concurrently elevated in the clinical range. Not only are these dimensions—often classified into externalizing (behavioral excess) or internalizing (behavioral deficit) dimensions—problematic concurrently, the externalizing and internalizing diagnoses are often found to be comorbid. For example, adolescents hospitalized on a psychiatric unit are often described by more than one diagnosis or by a diagnosis that reflects different dimensions of psychopathology. It is not unusual for such adolescents to demonstrate significant levels of depression or anxiety and at the same time demonstrate pathologies of behavioral excess (oppositional defiant disorder, conduct disorder, attention deficit hyperactivity disorder (ADHD); see, for example, O'Connor, McGuire, Reiss, Hetherington, & Plomin, 1998). The most useful assessment procedure will provide a quantified estimate of the pattern of problems in need of remediation and answer relevant questions such as the role that family conflict may play in symptom development or maintenance, and the context of such symptoms.

Multidimensional assessment is uniquely designed to describe patterns of adjustment characterized by dimensions of both poor adjustment and relatively symptom-free adjustment. In many cases the documented absence of certain problems is as important as is the documented presence of others. The following common clinical challenge documents the

importance of multidimensional assessment. Children and adolescents are often referred for evaluation by parents, educators, physicians, and mental health professionals because they have been observed to be hyperactive and/or inattentive. Exclusive application of homogeneous measures of ADHD or hyperactivity in such cases can only prove to be problematic. These referred children are often found to be "inattentive" for reasons other than ADHD (depression, anxiety, situational adjustment, learning disability, acquired cognitive disability, etc.). When significant elevation is not obtained on focused hyperactivity scales, guidance in the search for likely alternative diagnoses is not provided. On the other hand, youth found to meet all of the criteria of ADHD often present with coexisting conditions (see, for example, August, Realmuto, MacDonald, Nugent, & Crosby, 1996; Vaughn, Riccio, Hynd, & Hall, 1997). When such comorbid conditions remain unrecognized, their presence cannot be considered in treatment planning. In such cases, the resulting treatment may be compromised by the absence of such information.

4.13.1.3 Multisource Assessment

Although the evaluation of adjustment or psychopathology in adults usually relies on self-report, such self-description is rarely adequate in the evaluation of children and adolescents. Indeed, the context of assessment is fundamentally different for children and adolescents who are unlikely to refer themselves for evaluation or treatment and may not possess the academic, cognitive, or motivational status to complete a comprehensive self-report instrument. The first consideration in this regard is that young elementary school children, perhaps students in K through 3, may be unable to describe themselves adequately through response to questionnaire statements. These children are unlikely to have mastered the range of vocabulary necessary to adequately describe dimensions of adjustment; such language competence is not usually attained before the fourth or fifth grade.

Another consideration is the reality that youth are most often referred for evaluation because they are either noncompliant with the requests of the significant adults in their lives or exhibit problems in academic achievement, often presenting with inadequate reading skills. It is therefore not unusual that completion of a self-report inventory of several hundred items by a high school student (e.g., the Minnesota Multiphasic Personality Inventory-Adolescent consisting of 478 statements) could present quite an assessment challenge.

Parents and teachers not only refer youth for assessment, they are also the primary sources of useful systematic observation. Certainly adults should be the ideal informants to report noncompliance of a child to their requests. Parents are the only consistently available source for report of early childhood development and description of child behavior in the home. Teachers offer the most accurate observations of the age-appropriateness of a child's adjustment in the classroom and academic achievement, as well as of the attentional, motivational, and social phenomena unique to the classroom and to the school. It is likely, however, that such observational accuracy decreases after the elementary school grades, as teachers have very little continuous observation of students who appear for only 45 minutes a day in a classroom of 30 or more other students. (An exception perhaps would be the ratings of counselors, resource room teachers who work with small groups of students, and teachers in self-contained special education classrooms.) Youth self-description, regardless of problems that have been documented for this source of information (Greenbaum, Dedrick, Prange, & Friedman, 1994; Jensen et al., 1996), still must be seen as the most direct and accurate expression of personal thoughts and feelings. (Note that Michael & Merrell (1998) have demonstrated adequate short-term temporal stability for the self-report of third to fifth graders.)

The availability of a coordinated "family of instruments" in which uniquely designed questionnaires are completed by two or three different informant sources, each providing independent sets of child descriptions, offers a natural opportunity for critical comparison of the value of these sources of information. Achenbach, McConaughy, and Howell (1987) conducted a comprehensive literature review and found, although relatively greater between-source agreement was obtained for scales representing externalizing behaviors, very limited concordance in general between the report of parent, teacher, and youth. A review of similar studies that evaluated the responses to objective interviews from parent and child concluded that greater agreement was obtained for visible behaviors and for child-parent pairs with increasing child age (Lachar & Gruber, 1993).

Although one reasonable approach to the interpretation of differences between parent, teacher, and youth is to assign such differences to situation-specific variation (e.g., the child is actually oppositional at home, not in the classroom), other explanations are equally plausible. One alternative explanation of cross-informant variance requires a close examination of the measures being applied. It is a frequent

occurrence to find that scales with similar names contain significantly different content. In this situation an apparent lack of agreement between informants on similarly named scales (such as “depression”) may more accurately reflect the differences in statement manifest content. At one extreme a clinician might consider accepting any evidence of symptom presence from any informant source, at the other extreme it would also be possible to focus exclusively on problems demonstrated by at least two or even all three informant sources.

On the other hand, the development and application of valid identical across-source measures may not provide optimal assessment data. Such an approach may restrict the diagnostic potential of each informant source by excluding the measurement of attributes that may be uniquely obtained from only one informant source (e.g., a parent-completed measure of developmental delay, a teacher-completed measure of academic skills, or a youth-completed measure of achievement motivation).

Another explanation for poor cross-informant agreement for data gathered in a clinical context acknowledges the potential effect of response sets on the accuracy of such information. Starting with the child or adolescent being assessed, adequate compliance with questionnaire instructions may not be obtained—reflecting either inadequate language or reading skills, or lack of adequate motivation for the task. It is also likely that a youth may not want to share a personal history of maladaptive behavior and current internal discomfort with mental health professionals, although a negative presentation of parent adjustment and home conflict may be more readily provided. At times, youth may also be motivated to admit to problems and symptoms that are not present. These same motivations and conditions may also influence parent report. Indeed, there has been some concern that poor parent adjustment may compromise the validity of parent report (Achenbach, 1981). However, a subsequent review (Richters, 1992) and the specific analysis of this issue with the Personality Inventory for Children (PIC) (Lachar, Kline, & Gdowski, 1987) have found no empirical support for this consideration. Some instruments, however, incorporate validity scales to identify such response sets. These scales are designed to measure random or inadequate responses to scale statements, defensive denial of existing problems, as well as admission of symptoms that are unlikely to be present or exaggeration of actual problems. (Examples of the effect of informant defensiveness on profile pairs are presented in Table 5 and discussed in associated text for the Personality Inventory for Youth (PIY) and the Personality

Inventory for Children, 2nd Edition (PIC-2). See also Lachar, in press.)

For the clinician there is a distinct pragmatic advantage in using an assessment system with comparable parent, teacher, and youth versions. Research may document one day that one source of descriptive report is superior to another in the assessment of a specific problem domain (only surveys of the opinions of clinicians and parents are currently available: Loeber, Green, & Lahey, 1990; Phares, 1997). Until that day, however, conditions regularly occur in the conduct of psychological evaluations that on some occasions make it necessary to collect one, and on other occasions another, of these parent-, teacher-, and youth-report questionnaires. Children may be too young, uncooperative, or language-impaired to complete a questionnaire. The evaluation may occur during the summer vacation, or the youth may have not consistently attended one school, or may have left school permanently, making collection of a teacher rating difficult or impossible. Parents may fail family appointments when their child is hospitalized, or children may be under agency guardianship rather than parent supervision. In such not infrequent cases, a “family” of objective questionnaires separately normed for parents, teachers, and youth provide the flexibility to facilitate the collection of data from the source or sources that are available.

4.13.1.4 Psychometric and Technical Considerations

The topics discussed in this section are unlikely to spark much enthusiasm in clinicians who also demonstrate little interest in reading the technical sections of test manuals. These issues, however, are important as they influence test validity and ease of interpretation and must be considered in the selection of instruments for specific clinical applications.

4.13.1.4.1 Scale stability and homogeneity

It may seem quite reasonable to apply the technical standards and statistical techniques first developed in the construction of ability and achievement measures to scales of adjustment and personality. For example, in the development of an age-normed measure of verbal IQ or a grade-normed measure of arithmetic calculation skill, certain assumptions are made. As each of these measures would be conceptualized as relatively homogenous and stable over at least weeks or months, item components would need to demonstrate substantial intercorrelation as well as little change over time (coefficient alphas

and test–retest correlations of one week/three months of at least 0.80).

Measures of adjustment or psychopathology, in contrast, can vary in degree of homogeneity and may in fact in some cases be expected to be relatively unstable over time or sensitive to the effects of treatment. A measure of trait adjustment might consist of responses that describe previous behaviors and therefore demonstrate considerable temporal stability, while a measure of mood that consists of self-statements made in the present tense may provide a state measure that is expected to vary considerably over time. An empirically constructed scale in which items have been selected because they have statistically separated well-defined samples (e.g., normal adolescents vs. adolescents adjudicated delinquent) is likely to be multidimensional and limited in homogeneity. In contrast, a scale constructed by factor analysis consists of items that correlate highly with each other and such scales demonstrate substantial homogeneity. The importance of scale homogeneity as measured by coefficient alpha and scale stability represented by the test–retest correlation must be evaluated in light of the expressed meaning and application of each measure. Of importance in any case is the demonstrated validity of these scales; that is, evidence that each scale measures what it is supposed to and that interpretive guidelines are derived from empirically based relationships between scale values and independent observations.

4.13.1.4.2 Scale construction methods and scale character

It is important to be aware of both scale content and scale performance when applying scale standard scores in clinical assessment. Scale content has gained in judged importance and scale items have been selected frequently because of specific content or written to measure specific dimensions (see, for example, the Conners *DSM-IV* scales (Conners, 1997) and the Disruptive Behavior *DSM-IV* scales of the Student Behavior Survey (Lachar, Kline, Wingenfeld, & Gruber, 1999; Pisecco, Lachar, Gallen, Gruber, & Huzinec, 1998). This recent focus on item content is in sharp contrast to previously developed empirically keyed scales in which item performance took preference over item content. Current consensus is that scale content is necessary but insufficient to assure scale validity. It is recommended that clinicians first read the items on each scale and examine the content dimensions provided as well as the content equivalence for parent, teacher, and youth questionnaires before attempting any clinical application.

It is of course generally true that increased scale length will improve the depth or variety of expressed content. Against this reality must be balanced concerns about the time required of informants to complete a questionnaire, an issue that appears to be significant to psychologists who work in test development and those who work in test application. One must also consider the context in which evaluations are to be conducted. A baseline or initial evaluation would justify a more comprehensive evaluation than would a repeated evaluation or some approach to treatment evaluation in which daily or weekly observations are required.

Each clinician will need to judge the value of the information obtained vs. the time needed to generate these measures. In this author's experience, within a clinical (vs. research) context, parents view these questionnaires as appropriate and appreciate such an offer to make this sort of contribution to the diagnostic process. Teachers, unless asked to complete lengthy forms on a whole classroom of students, are quite willing to spend 15 or 20 minutes completing a form that either initiates a referral to a school psychologist or provides necessary classroom observation for an independent or agency-conducted evaluation. The instruments of the families of questionnaires presented in this chapter vary in length from a "short form" of 10 items that generates one standard score to a "short form" of eight 12-item shortened scales that is as long as the full-length versions of some other questionnaires.

Scale length (number of items) and rate of item endorsement (infrequently endorsed items in short scales) within normative and clinical samples directly influence scale performance, such as range of standard scores and the shape of score distributions. Response format is also relevant in that a scale score increment of one point may represent the endorsement of one additional symptom or characteristic (true/false; present/absent) or a subjective increase in symptom observed certainty or frequency (somewhat true to very true; sometimes true to often true). It may be useful to identify the minimum descriptive content necessary to generate a clinically significant standard score. It may also be useful to determine the amount of response difference necessary to move a scale value from normal limits to the clinical range. Note that a scale with a 0–18 raw score range can be generated from either six symptoms rated from 0 to 3, or from 18 true/false statements that vary in both symptom intensity and frequency. Although it will take longer to respond to 18 than to six items, it is obvious that the content basis for standard scores will vary substantially for these two measures. A comprehensive

evaluation of any dimension is likely to require more than a handful of observations.

Although these scale scores are usually positively skewed (symptoms and dimensions of psychopathology are not normally distributed), scale distribution may be excessively restricted due to scale length, item response frequency, or normative sample character so that the scale mean and standard deviation that generate standard scores are very limited. (See Rowe & Rowe (1997) for a contemporary discussion of this issue with a focus on the Conners' Global Index.) For example, a scale with a normative mean of 1.13 and a standard deviation of 2.78 (Conners' Teacher Rating Scale-Revised (CTRS-R): Oppositional, female 12–14 years) would generate the following raw to *T*-score conversions: 0 = 46*T*, 1 = 50*T*, 2 = 53*T*, 3 = 57*T*, 4 = 60*T*, 5 = 64*T*, 6 = 69*T*. Because the Oppositional scale consists of six items rated from 0 (not true at all/never, seldom) to 3 (very much true/very often, very frequent), the following interpretive issues are suggested. Considering that the scale raw scores for Oppositional can range from 0 to 18, responses of "not true" or "just a little true" predominate in the normative sample. Only a scale score of "0" extends below the mean. Of most interest, a clinically elevated score (> 65*T*) can be obtained by responding "just a little true" to each item, or "very much true" to only two statements. Please note that other CTRS-R scales actually obtain mean raw scores below 1.0 and standard deviations below 2.0, but the CTRS-R is far from the only questionnaire to exhibit these troublesome psychometric characteristics. Users are well advised to review a test's raw score-to-*T* score translation table or profile where the issue can be evaluated.

4.13.1.4.3 Evidence of scale and questionnaire validity

All too often test evaluations and reviews discuss internal consistency and reliability at length and then give short shrift to test validity. This is perhaps understandable in that it is easier to make comparisons and establish numerical guidelines for evaluating these characteristics. Unfortunately—and particularly so for instruments rating psychopathology or other infrequent behavior—the emphasis should be quite the opposite. Validity is far more difficult to define, measure, and establish, but it really is the only bottom line. Even a scale that demonstrates somewhat weak or even impaired reliability can be extremely valuable clinically if it can successfully identify a rare or transient characteristic such as violent labile aggression. It is necessary to establish the validity of specific

applications of each scale before it can be interpreted with certainty. It should be understood that standard scores derived from the most representative and comprehensive of normative data simply provide the frequency or infrequency of a given score in this normative sample. Just because a pattern of scale standard scores is infrequent does not necessarily demonstrate that this pattern is clinically meaningful or significant. It is also important to acknowledge that scale validity is not a general attribute. The validity of a scale can only be demonstrated within a specific context for a specific purpose.

The construction or selection of a normative sample for each questionnaire is also a process requiring critical review—especially when questionnaires provide a variety of normative options (see, for example, Reynolds & Kamphaus, 1992). In the evaluation of school-aged children, scale raw scores may be converted to standard scores based upon normative (not receiving counseling or special education services), representative (incorporated an accurate proportion of all children), or referred (clinical or special education) samples. Normative samples also differ as to whether they are gender specific or combine the protocols of both boys and girls. It should also be noted that normative samples may be constructed to reflect child age. Such age-referenced samples may represent broadly defined groups such as preadolescent/adolescent, or provide norms that reflect a much more narrow age range (see Conners, 1997). The relative utility of any normative sample rests upon the demonstration that derived standard scores are comparable in validity. For example, the scales of the PIC have been found to be equivalent in their ability to predict external criteria for boys vs. girls and for children vs. adolescents (Kline & Lachar, 1992; Kline, Lachar, & Sprague, 1985). Test authors who promote the use of an atypical normative procedure should demonstrate that the resulting interpretive guidelines are as accurate as those obtained by more traditional means. For example, if a questionnaire provides both traditional gender-specific and novel mixed-gender norms, evidence should be provided that mixed-gender norms generate interpretive guidelines which are either more accurate or as accurate as more traditional gender-specific guidelines.

The construction of a new set of scales for a multidimensional inventory or set of inventories must follow a path in which fundamental evidence of validity must first be obtained. If obtained, additional study of more specific and focused evidence of accurate scale performance must be pursued. For example, when items are placed on dimensions on the basis of manifest

content, the accuracy of this placement could be demonstrated through item-to-total correlations in which each item will be expected to obtain the greatest correlation with the dimension on which it has been placed. If the scales are constructed to measure adjustment or psychopathology, each scale could then be applied to regular education and special education and/or clinical samples to determine if meaningful separation occurs between these groups for each scale. Here “meaningful” is more than a statistically significant difference, which when samples are large could be so small as to be of no pragmatic value. Scales in this context should demonstrate at least a moderate effect size—generally about one-third of a standard deviation or greater (Cohen, 1988).

Unlike ability tests that gain initial construct validity by correlation with similar although established measures, scales of adjustment are less likely to demonstrate comparable concordance (0.70s–0.80s) because of differences in content and format. Indeed, substantial correlations obtained between pairs of scales from different tests completed by the same informant most likely reflect substantial shared item content. Adjustment scales for children and adolescents acquire validity through demonstration of external correlates that may take a variety of forms including independent ratings, special education placement, diagnoses, or specific treatment. For example, in students referred for evaluation, elevation of a parent-completed attention deficit hyperactivity scale should correlate with content-congruent teacher ratings, and students diagnosed with ADHD and assigned to receive stimulant therapy should obtain clinically elevated scale scores that subsequently fall within the normative range following successful treatment.

The accumulation of substantial construct validity and the development of empirically based interpretive guidelines are prerequisite for a questionnaire to achieve status as an established clinical measure. Unfortunately, the accumulation of a substantial bibliography provides an insufficient basis for the development of this status. Tests are far more often applied in the quantification of phenomena under study rather than applied in the study of scale score meaning. For example, contrast “How many children with mild mental retardation also demonstrate the *DSM-IV* criteria for a disruptive behavior disorder?” with “How do children with mild mental retardation who also obtain an elevated parent-informant hyperactivity scale differ in some way from children with mild mental retardation who do not obtain such an elevated scale?” That is, tests of adjustment are usually applied in the paradigm

in which the investigator asks the question “What is the likelihood of a given test result given certain clinical criteria?” vs. “What is the likelihood of certain clinical phenomena given a certain test score, score range, or score pattern?” It is the latter question (and associated research design) that replicates the use of tests in clinical assessment and hence offers the better evidence of validity.

The clinician who reads that 95% of a sample of adolescents hospitalized on a psychiatric unit with a discharge diagnosis of major depression obtain an elevated score on a certain depression scale should understand that this information is insufficient to demonstrate either construct validity or provide an interpretive guideline. If, for example, a comparable proportion of the remaining adolescents on that psychiatric unit also received similar test results, these additional data would suggest that this “depression” scale measures general distress rather than depression. In addition, clinicians should pay attention to any evidence of a scale’s ability to make clinically meaningful distinctions, not the separation of normal and referred subjects (see, for example, Forbes (1985) in which PIC profiles of children with the diagnosis of ADHD were compared to the PIC profiles of behaviorally disturbed children whose symptom presentation did not support such a diagnosis). It is also important to focus on studies that document the independent correlates of scale elevations or profile patterns (see, for example, LaCombe, Kline, Lachar, Butkus, & Hillman (1991) and the clinician, teacher, and self-report correlates of parent-report PIC-2 shortened scales presented in Table 4). In general, such studies assemble a sizable data set that includes both the test scores under study and other clinically meaningful independent information, with the goal of determining the degree to which test variables accurately predict clinically meaningful variables across all subjects. In this way a scale’s convergent validity (scale elevation for patients afflicted with the problem to which the scale is directed) and discriminant validity (scale nonelevation for nonafflicted subjects or patients afflicted with unrelated problems) (Campbell & Fiske, 1959) are established.

4.13.2 MULTIDIMENSIONAL MULTISOURCE RATING SCALES

4.13.2.1 Overview

The second section of this chapter describes the four families of commercially available multidimensional, multisource questionnaires. Although considerable detail is provided for each set of scales, clinicians are urged to review

all materials that may be appropriate for their specific applications. The conduct of this review has been difficult. Within these four sets of scales, one (Behavior Assessment System for Children; Reynolds & Kamphaus, 1992) was first published in 1992, two (Child Behavior Checklist and related measures: Achenbach, 1991; Conners Rating Scales (CRS): Conners, 1997) have recently been revised and expanded, while the fourth set of scales includes one recently published questionnaire (Personality Inventory for Youth: Lachar & Gruber, 1995), one questionnaire just revised (Personality Inventory for Children, 2nd Edition (PIC-2): Lachar, 1999), and one new measure (Student Behavior Survey (SBS): Lachar, Kline, Wingensfeld, & Gruber, 1999). Although the established materials have been collectively applied in approximately 3000 studies, no systematic review has been conducted to identify those studies that could form the basis for the development of interpretive guidelines (note in contrast the actuarial studies of profile scales and profile types for the PIC: Lachar & Kline, 1994).

4.13.2.2 Conners' Rating Scales-Revised

[Multi-Health Systems Inc., 908 Niagara Falls Boulevard, North Tonawanda, NY 14120-2060, USA]

The Conners scales were derived from teacher and parent measures first used in research conducted during the 1960s in the study of the pharmacological treatment of disruptive behaviors. The Conners' Rating Scale-Revised (CRS-R) emerged from a considerable research base generated over more than 30 years. An annotated bibliography of over 450 studies is available from the publisher (Wainwright & MHS Staff, 1996). Parent, teacher, and adolescent response to questionnaire items takes the same four-choice options: 0 = Not True at All (Never, Seldom), 1 = Just a Little True (Occasionally), 2 = Pretty Much True (Often, Quite a Bit), and 3 = Very Much True (Very Often, Very Frequent). CRS-R standard scores are linear *T* scores derived from contiguous three-year segments of the normative sample. The CRS-R hand scoring forms are easily completed and scored, and computer support for scoring and interpretation is available. The 1997 manual (Conners, 1997) and most current review chapter (Conners, in press) suggest that these scales continue to focus on ADHD and strengthen their assessment of related or comorbid disorders.

Conners in these publications emphasizes the relative importance of teacher observation in making these diagnostic determinations and

suggests that the comparison of sources (teacher, parent, adolescent) can determine the reliability of the generated diagnostic hypotheses. Conners also suggests that the CRS-R scales should be augmented when their internalizing scales are elevated by using other measures, such as the Children's Depression Inventory (CDI) (Kovacs, 1992) and the Revised Children's Manifest Anxiety Scale (Reynolds & Richmond, 1985).

Conners details his concern for the accuracy of each protocol and the possible effect of response sets, such as defensiveness, although the CRS-R does not provide measures of such response sets. Conners suggests that it is useful to review the within-scale consistency of the content of items rated "3" (very much true/very often, very frequent) within elevated ($>64T$) scales/indices, although the efficiency of this process would be vastly improved through the development of a consistency scale. Although the manual suggests that a clinically elevated scale exceeds $64T$, case examples present scale elevations below $60T$ as contributing to the diagnostic process. The manual presents and analyzes raw scores (mean, standard deviation) rather than providing *T*-score ranges and *T*-score frequencies within well-defined clinical or special education samples (e.g., establishing the scale *T*-score ranges for the dimensions that best define children diagnosed with ADHD).

4.13.2.2.1 Conners' Parent Rating Scale-Revised

The 80-statement form is the most comprehensive of the Conners' parent measures. Although the manual notes the reading level of this form as ninth grade, Conners' Parent Rating Scale-Revised (CPRS-R) items range in complexity from one-word stems such as "Irritable" to a 28-word descriptive phrase: "Does not follow through on instructions and fails to finish school work, chores or duties in the workplace (not due to oppositional behaviour or failure to understand instructions)." The CPRS-R generates seven factor-derived nonoverlapping scales apparently generated from the ratings of regular education students from the normative sample ages 3–17 years (Oppositional, Cognitive Problems, Hyperactivity, Anxious-Shy, Perfectionism, social problems, and psychosomatic), Conners' ADHD Index, Conners' Global Index, and *DSM-IV* scales Inattentive, Hyperactive-Impulsive, and Total. The ADHD Index consists of the 12 items empirically determined to best identify children at risk for an ADHD diagnosis (this 12-item set varies by informant source across this family of tests). The Conners' Global Index includes the

10 items historically labeled the “Hyperactivity Index” in previous published and unpublished versions of the Conners parent and teacher scales. This index is presented as the best general measure of behavioral adjustment and the one most sensitive to treatment effects.

Three additional forms are derived from these 80 items: (i) a shortened form of 27 items (CPRS-R:S) that generates the 12-statement ADHD Index and shortened forms of the Oppositional, Cognitive Problems, and Hyperactivity scales, (ii) a 26-item format that provides the ADHD Index and the three *DSM-IV* scales (CADS-P), and (iii) a sheet that only presents the 10-item Global Index. These three shortened forms are provided as options to monitor treatment effects.

Table 1 presents the pattern of scale and Index overlap (scales: Oppositional, 10 items; Anxious-Shy, 8 items; Perfectionism, 7 items; Social Problems, 5 items; and Psychosomatic, 6 items are each formed from unique items).

The *DSM-IV* Hyperactive-Impulsive and Hyperactivity scales exhibit considerable overlap as do the *DSM-IV* Inattentive and Cognitive Problems scales, suggesting the need for additional guidelines for application and interpretation of the full-length and shortened parent and teacher scales. It is also interesting to note that the Global Index, presented as the 10 items most sensitive to behavioral disturbance, almost never appears on any other scales (eight out of 10 items are unique).

4.13.2.2.2 Conners' Teacher Rating Scale-Revised

The full-length version of the Conners' Teacher Rating Scale-Revised (CTRS-R) consists of 59 items. Across-informant analysis may be facilitated in that the six CTRS-R factor-derived scales from three- to 17-year old normative children have been given the same names as the CPRS-R scales (there is not a teacher version of the CPRS-R Psychosomatic scale). The Conners' Global Index, Conners' ADHD Index, and three *DSM-IV* scales of the

CPRS-R are also incorporated in this form. Three short teacher forms are also available: (i) a 28-item version that provides shortened Oppositional and Cognitive Problems scales, the Hyperactivity scale, and the ADHD Index, (ii) a 27-item version that provides the 12-item ADHD Index and the three *DSM-IV* scales, and (iii) a single sheet that presents the 10-item Global Index.

A comparison of CPRS-R and CTRS-R scales and index item content and scale length may shed some light on CRS-R construction and suggest issues in clinical application. Because the CTRS-R is shorter (59 items) than the CPRS-R (80 items), CTRS-R factor-derived scales are also shorter, with individual responses contributing greater to score derivation. The greatest shortening is present on factor scales Oppositional (10 to six items) and Cognitive Problems (12 to eight items). Assignment of identical titles to CPRS-R and CTRS-R scales should not lead the clinician necessarily to assume comparable item content for parent-teacher scale pairs. While the Global Index and *DSM-IV* scales do consist of items with identical or comparable content, the remaining CTRS-R score dimensions, except for the Cognitive Problems and Hyperactivity scales, consist of no more than 50% comparable content for similarly titled CPRS-R dimensions (six of the eight CTRS Cognitive Problems items have equivalents on the identically named 12-item CPRS dimension. Six of the seven CTRS Hyperactivity items have equivalents on the identically named nine-item CPRS dimension). Therefore, it was not unexpected to observe that the greatest parent-teacher agreement for normative cases occurred on the Global Index and *DSM-IV* measures (Connors, 1997, p. 126).

4.13.2.2.3 Conners-Wells' Adolescent Self-Report Scale

This 87-item questionnaire has been designed for students ages 12–17 with at least a sixth grade reading competence. Conner-Wells'

Table 1 CPRS-R item overlap.

Measure (no. of items/no. of unique items)	B	C	H	K	L	M
B:Cognitive problems (12/6)			3		5	
C:Hyperactivity (9/3)				1		5
H:Conners' ADHD Index (12/6)	3			1	2	2
K:Conners' Global Index (10/8)		1	1			
L: <i>DSM-IV</i> Inattentive (9/4)	5		2			
M: <i>DSM-IV</i> Hyperactive-Impulsive (9/2)		5	2			

Adolescent Self-Report Scale (CASS) includes six nonoverlapping factor-derived scales with lengths of either eight terms (Anger Control Problems; Hyperactivity) or 12 items (Family Problems; Emotional Problems, Conduct Problems, Cognitive Problems). This form can also be scored for a 12-item ADHD Index and two *DSM-IV* ADHD scales (as in the parent and teacher versions). Two shortened versions are available. In one, 27 items yield shortened versions of three factor-derived scales (Conduct Problems, Cognitive Problems, and Hyperactivity) and the 12-item ADHD Index, while the other provides the ADHD Index and the two *DSM-IV* scales Inattentive and Hyperactive-Impulsive in 30 items.

Comparison of item content between adolescent and parent report demonstrates excellent concordance for *DSM-IV* scales. Although reasonable agreement would be expected to be found for ADHD Indexes, Cognitive Problems and Hyperactivity scales, and between CPRS-S Oppositional and CASS Anger Control Problems, very few items from the parent scale can be identified in the self-report version.

4.13.2.2.4 Commentary

Table 2 displays average *T*-score group estimates derived from mean raw scores for parent, teacher, and self-report scales from three samples: general clinical, ADHD, and matched normative (Conners, 1997, pp. 135–136).

It would be very useful for the CRS-R Manual to provide for each sample the proportion of scale scores that equaled or exceeded the minimum *T*-score for the clinical range, whether this is 60*T* or 65*T*. Such lack of clinically relevant information is also demonstrated in a journal presentation of the six factor-derived CASS scales (Conners et al., 1997). In comparing 86 adolescents with a sole diagnosis of ADHD combined type to matched normative controls, an overall correct classification rate of 82.6% was obtained using all six scales. Although all six scales were statistically significant in group contrasts (demonstrating comorbidity and need for multidimensional assessment), no indication of the relative importance of each scale's contribution to this discrimination is given, nor are the proportion of elevated scores per scale provided for ADHD and contrast samples.

Table 2 documents substantial normative/clinical differences as well as the relative superiority of parent report when it is compared to the companion self-report form. Brevity and ease of application would recommend the CRS-R measures in monitoring the treatment of children with ADHD. Versions of *DSM-IV* scales for parent, teacher, and youth will allow

the study of the accuracy of each informant source in diagnosing ADHD as well as the study of across-informant agreement.

The majority of CRS-R measures appear to have been developed out of distinctly different item pools, resulting in sets of scales that may complement rather than duplicate each other. It would be useful for the manual to explain why the CRS norms incorporate three-year intervals. The current profiles suggest in displayed raw-to-*T*-score conversion substantial and variable gender and age effects. The magnitude of this age effect is easily demonstrated by tracking on a profile the *T*-score equivalents for one scale raw score across the five sets of age norms. For example, a *DSM-IV* Hyperactive Impulsive raw score of 12 for males resulted in the following five *T* scores: 3–5, 60*T*; 6–8, 63*T*; 9–11, 66*T*; 12–14, 70*T*; 15–17, 81*T*. In contrast, the ADHD Index demonstrated far less age-related variation.

4.13.2.3 Child Behavior Checklist/4–18 and 1991 Profile (CBCL); Teacher's Report Form and 1991 Profile (TRF); Youth Self-report and 1991 Profile (YSR)

[University Associates in Psychiatry, 1 South Prospect Street, Burlington, VT 05401-3456, USA]

The CBCL, the first published measure of these three (other forms now cover preschool, direct observation, and young adults), was initially published in 1983 as a direct extension of the factor analytic study of child problems and symptoms first published in 1966 (Achenbach, 1966). The 1983 CBCL consisted of two parts, an assessment of competence and a series of factor-derived behavior problem scales constructed from 120 brief descriptions. Each checklist item is rated by a parent as either 0 = Not True (as far as you know), 1 = Somewhat or Sometimes True, or 2 = Very True or Often True, with several items requiring individual elaboration when these items are positively endorsed (scores of 1 or 2). Because of these 17–21 items in each checklist, the informant must be closely monitored and an experienced clinician must review these comments before each pamphlet is submitted to more automatic scoring procedures (Drotar, Stein, & Perrin, 1995). A companion teacher rating form became available in 1986 and a self-report checklist was published in 1987. These forms and subsequent revisions have been used in over 2000 citations (Vignoe & Achenbach, 1997). The popularity of the CBCL and related instruments in research application has had a significant influence on the study of child and adolescent psychopathology since 1986.

Table 2 Average approximate *T*-scores for ADHD, clinical (CLIN), and normative (NORM) samples.

	<i>ADHD</i>	<i>CLIN</i>	<i>NORM</i>
<i>CPRS-R</i>			
Oppositional	59	63	47
Cognitive Problems	67	57	46
Hyperactivity	69	58	48
Anxious-Shy	56	58	49
Perfectionistic	49	57	50
Social Problems	63	68	48
Psychosomatic	56	57	48
Global Index	67	64	46
ADHD Index	66	59	47
<i>DSM-IV</i> Total	69	59	47
<i>CTRS-R</i>			
Oppositional	61	64	47
Cognitive Problems	56	50	47
Hyperactivity	69	68	49
Anxious-Shy	62	62	46
Perfectionistic	56	62	49
Social Problems	51	58	53
Global Index	66	62	46
ADHD Index	65	58	48
<i>DSM-IV</i> Total	64	58	46
<i>CASS</i>			
Family Problems	53	53	46
Emotional Problems	56	56	47
Conduct Problems	57	54	45
Cognitive Problems	60	56	44
Anger Control Problems	57	57	47
Hyperactivity	57	52	45
ADHD Index	59	53	44

In each of these three forms competence has been measured through rationally developed scales that vary by informant source. The factor-derived dimensions of child problems and symptoms followed a similar general structure for each informant source (each designated an "axis" by Achenbach). Each form's profile recorded several narrow-band scales (designated "syndromes" by Achenbach), summative Internalizing and Externalizing scales derived from second-order factor analysis, and a Total Symptoms scale. A series of item analyses (principal components analysis followed by varimax rotation) was applied to independent clinical samples of boys or girls within several age-range defined samples, resulting in dimensions often unique for one gender or one age group, as well as scales with the same titles but differing in content. The products of what appears to be a direct (vs. guided) unrepeated application of a data reduction technique of scale construction proved to be problematic. These item analyses apparently identified at times either sample-specific or age-/gender-specific variance rather

than broadly demonstrable dimensions. For example, only the profile for 12–16-year-old boys did not include a depression dimension. (Certainly this result could not be taken as support for the conclusion that adolescent boys do not demonstrate problematic depression!) Obvious problems presented themselves in applying these scales in research and clinical assessment. Repeated administration to the individual case over time could result in changes in scores as likely to be related to instrument or normative variation as to actual changes in subject status.

The 1991 revision, documented in five monographs totaling over 1000 pages, represented a major departure (Achenbach, 1991a, 1991b, 1991c, 1991d, 1993). This revision, in contrast to the earlier versions of these checklists, emphasized consistencies in scale dimensions and scale content across child age (4–18 years), gender, and respondent/setting. Because the actual checklist items remained essentially the same, this revision also provides substantial continuity, allowing either the continued use of the original sex-age-specific narrow-band

scales, or the rescaling of previously obtained protocols to generate the new scales. (Readers may find it useful to first read Chapter 3 of Achenbach (1993) to gain an overall appreciation of these checklists.)

Of primary importance is that a series of within-instrument item analyses was conducted to obtain dimensions common across gender and age groups ("core syndromes"). Substantial samples of protocols from each form obtained from clinical and special education settings were evaluated (4455 CBCLs, 2815 TRFs, and 1272 YSRs). This current version also provides eight narrow-band scales for each informant derived from analysis of the 89 items common to parent-, teacher-, and self-report forms, as well as Internalizing, Externalizing, and Total Scales. The CBCL Internalizing scale score is derived from the responses of three narrow-band scales: Withdrawn, Somatic Complaints, and Anxious/Depressed. The Externalizing scale score includes the responses from two narrow-band scales: Delinquent Behavior and Aggressive Behavior. Three other common scales were not placed on these two broad-band dimensions: Social Problems, Thought Problems, and Attention Problems. Review of scale content reveals some item overlap across these eight scales within each form, substantial difference in scale length within a profile (e.g., TRF scales range from eight to 25 items), and scales that are identically named yet vary in content across informants. Close examination reveals that only one of these narrow-band scales ("cross-informant syndromes"), Somatic Complaints, presents the same nine items to parent, teacher, and youth.

The 1991 forms now rely on national vs. regional normative (nonreferred) samples. Although the CBCL and YSR are routinely self-administered in clinical application, the CBCL normative data and some undefined proportion of the YSR norms were obtained through interview of the appropriate informant. It may be that the interactive process of interview administration of the CBCL inhibited informant response to checklist items. For example, six of eight parent informant scales obtained average normative raw scores of less than 2 with associated restricted scale variance. It is important to note that increased problem behavior scale elevation reflects increased problems, although these scales do not consistently extend below 50*T*. Because of the idiosyncratic manner in which *T* scores are assigned to scale raw scores, it is useful to follow each manual's suggestion to use scale raw scores, rather than *T* scores, in research applications. It is difficult to determine the interpretive meaning of checklist *T* scores, the

derivation of which has been of concern (Lachar, 1993; Kamphaus & Frick, 1996; Kline, 1994, 1995). Scale values under 70*T* are normalized and therefore compressed and unrepresentative of actual scale score distributions in the normative sample. Values above 70*T* have been arbitrarily assigned by the checklists' author with the goal of placing all possible scale scores evenly across available profile space and do not represent any specific probabilistic or statistical relationship with scores under 70*T*. The CBCL checklists do not provide any measure of informant response set or protocol validity. Not only is the identification of either symptom denial or random response problematic, but the *T* score assignment process above 70*T* can result in protocols that are the result of a response set to exaggerate problems but appear as valid estimates of child adjustment.

4.13.2.3.1 *Child Behavior Checklist/4–18 (CBCL) and 1991 Profile*

Parents complete a four-sided pamphlet and clinicians complete one of two two-sided profile forms (one for girls, one for boys). Norms are provided for two age ranges (4–11 and 12–18 for problem scales, 6–11 and 12–18 for competence scales) separate by gender. Rated competence items are organized by manifest content into three narrow scales (Activities, Social, and School) which are then summed into a total score. Parents are asked to list and then rate (frequency, performance level) a child's participation in sports, hobbies, organizations, and chores. Parents also describe the child's friendships, social interactions, performance in academic subjects, need for special assistance in school, and history of retention in grade. As standard scores for these scales increase with demonstrated ability, a borderline range is suggested at 30–33*T* and the clinical range is designated at *T* < 30. It has been suggested that a child's social and economic opportunities as well as race and ethnicity significantly affect these values (Drotar et al., 1995). These scales have also been compared to those of the PIC in their ability to predict adaptive level as defined by the Vineland Adaptive Behavior Scales (Pearson & Lachar, 1994).

The behavior problem scales consist of one to eight word descriptions ("Overtired," "Argues a lot," "Feels others are out to get him/her"). Appendix A of the 1991 CBI provides scoring directions for the 21 items that incorporate informant individual comment. Clinicians must understand checklist content and child psychopathology to correctly score these items. For example, item 113 is "Please write in any

problems your child has that were not listed above.” The person who scores the checklist must review the up to three examples that can be elicited, determine if they duplicate other checklist content, and then score the one unique symptom that receives the highest rating (Very True or Often True rather than Somewhat or Sometimes True). The importance of this issue cannot be minimized as almost all items of the Thought Problems scale for all three informants are provided in this format. Syndrome *T* scores are interpreted as borderline at 67–70*T* and in the clinical range at *T* > 70. There is some evidence, however, which suggests that ADHD and comorbid conditions are more accurately identified with a lower cutting point (Biederman et al., 1993; Chen, Faraone, Biederman, & Tsuang, 1994; Faraone, Biederman, Weber, & Russell, 1998; Steingard, Biederman, Doyle, & Sprich-Buckminster, 1992). The Total Problem scale is interpreted as borderline at 60–63*T* and in the clinical range at *T* > 63. Although the summary Internalizing and Externalizing scales are discussed in terms of their relative elevation when compared to each other, absolute interpretive guidelines are not provided.

Profile and pamphlet formats facilitate review of response content at the item level. Although scales are developed to overcome the unreliability of item responses, CBCL content may be important for a number of reasons. On several scales a “very true or often true” response to as few as two or three statements places a scale *T* score in the clinical range. When scales are limited in length and relatively heterogeneous in content, the same scale elevation can be generated by quite different observational content. A casual review of scale content reveals some inconsistencies that follow into TRF and YSR dimensions. Obsessions and compulsions, often associated with anxiety disorders, join strange ideas, behaviors, and hallucinations on the Thought Problems dimension; “Nervous, highstrung, or tense” appears on Attention Problems as well as Anxious/Depressed; while “Unhappy, sad, or depressed” appears on Withdrawn as well as Anxious/Depressed. The 19–25 item Aggressive Behavior dimension appears to have a number of items that appear to be more readily associated with ADHD phenomena than any “aggressive” dimension: “Demands a lot of attention,” “Showing off or clowning,” “Talks too much,” as well as two items only on the TRF: “Talks out of turn,” and “Disrupts class discipline.” These observations suggest that although a factor analysis can identify clinically meaningful dimensions, subsequent efforts may be necessary to assure the stability of item scale placement. (See Erford (1996) for an example of an attempt to replicate

the structure of a short factor-derived teacher rating form.) It may also be useful to review the item by dimension correlation matrix that underlies each factor analysis to determine the appropriateness of each item’s scale placement.

4.13.2.3.2 Teacher’s Report Form (TRF) and 1991 Profile

The TRF emerged from the CBCL and the comparability of the 120 symptom/problem items is readily apparent, as are the eight syndrome, two broad-band, and one summary score. However, comparison of the competence sections of the TRF and the CBCL reveals a marked difference. TRF measures are based upon very limited data: an average rating of academic performance based on up to six subjects identified by the teacher, individual seven-point ratings on four topics (how hard working, behaving appropriately, amount learning, and how happy), as well as a summary score from these four items. The TRF designates a borderline interpretive range for the mean academic performance and the summary score of 40–37*T*, with clinical range of *T* < 37. Even a cursory review of the TRF pamphlet suggests that the equivalence of CBCL and TRF forms was given priority over collection of teacher observations that would be unique to the classroom and the school environment. If observations of social and academic skills were integrated into the TRF, the teacher could also provide estimates of adaptive behaviors necessary for success in the classroom.

4.13.2.3.3 Youth Self-report (YSR) and 1991 Profile

The elements of this self-report measure have most of the basic characteristics of the CBCL and TRF. The YSR requires a fifth grade reading ability and provides competence and problem behavior standard scores normed separately by gender for the age range of 11 through 18. Seven adaptive competency items are scored for Activities, Social, and a Total Competence scale. The YSR manual’s Appendix A is necessary to score these items (the addition of scoring examples would be very useful). These seven multipart items tap competence and level of involvement in sports, activities, organizations, jobs, and chores. Items also provide self-report of academic achievement, interpersonal adjustment, and level of socialization.

The remainder of the pamphlet presents 120 statements in which 16 are socially desirable items endorsed by most students. These items replace potential YSR equivalents of CBCL items that presented either problem behaviors

of very young children or school-related difficulties. The presence of these 16 unscored items represents an unused opportunity to develop a semantic inconsistency scale to measure random response or inadequate comprehension, or some sort of defensiveness measure. The lack of validity scales in this self-report of symptoms and problems is considered problematic by several reviewers (Kamphaus & Frick, 1996; Kline, 1995; Lachar, 1993). Two reviewers cite literature that documents the YSR's diagnostic insensitivity (Kamphaus & Frick, 1996; Merrell, 1994).

The eight problem scales range from 50 to 100 T , with 67–70 T assigned the borderline range and $T > 70$ the clinical range. The Activities and Social Competency scales range from 55 to 20 T , while the Total Competence score is profiled from 10 to 80 T . Activities and Social are in the borderline range from 33 to 30 T , while the clinical range includes scales $< 30T$. On the Total Competence scale, T scores of 40–37 are designated as borderline, while T scores < 37 form the clinical range.

4.13.2.3.4 Commentary

The CBCL, TRF, and YSR format is responsible for both the broad success that these instruments have achieved as well as their shortcomings. Each form can be completed in 10–20 minutes and appears easily scored (if competence items and items requiring informant explanation are avoided). Brevity is especially attractive for research applications in which subject and human subject research committee tolerance for multiple measures is of specific concern. Scoring software provided by the CBCL/TRF/YSR author facilitates not only the processing of individual instruments, but provides estimates of agreement over multiple reports on the same instrument, as well as agreement across different instruments.

It is almost certain that the CBCL, TRF, and YSR common format will stimulate meaningful across-informant research that will document the effect of item and scale content as well as informant characteristic (gender, age, ethnicity, clinical status) on both degree of agreement and relative contribution to the diagnostic process. Such research should generate useful guidelines for the clinical application of these instruments. On the other hand, the primary focus on parallel dimensions and consistency in the problem behavior scale construction technique limits collection of informant-specific observation. The other instrument families presented in this chapter vary in dimensions assessed to some degree to facilitate the collection of such information.

The decision to revise scales without modification of the original problem statement pool assured continuity but at the same time restricted potential improvement. These forms continue to demonstrate their original limitations in psychometric character. It is difficult to understand why the author has taken an idiosyncratic and statistically unsupportable approach both in subjecting summations of symptoms to a normalization process and in providing profiles that assign T scores to all physically possible raw scores. Certainly problematic item performance in the form of restricted range in symptom frequency within these normative samples is at the core of attempts to modify score distributions to improve psychometric performance. The absence of validity scales continue to present random, defensive, and exaggerated protocols as profiles that should be interpreted rather than excluded from this process.

The most remarkable observation, in contrast to the continuing research growth of CBCL/TRF/YSR application, is the apparent absence of interpretive guidelines for the clinician (certainly percentiles derived from normative samples do not provide evidence of scale meaning beyond that suggested by manifest item content). These 1991 manuals present as primary evidence of validity that items and scales differentiate clinical and normative samples. Such evidence is only clinically useful if such a decision is frequently made in practice. (Clinicians know whether a student has or has not been referred for an evaluation.) Certainly there should be sufficient opportunity within 2000 published studies to document the answers to specific clinically relevant questions. (Indeed, it is quite clear that much of the results of research in child psychopathology over the last 15–20 years have been dependent on the structure and psychometric characteristics of the CBCL.) It should be quite easy to identify and summarize those studies that used a contrasted-groups design to identify those scales that are effective in making specific distinctions.

A review of CBCL citations would easily find carefully selected samples relevant to clinical practice and scale validity. These protocols could then be scored for 1991 dimensions and subsequent profile types to establish the proportion of each sample that is elevated on each dimension or obtains a specific classification for each instrument. For example, one investigation would be presented in tabular form by profile scales that would be represented by 11 columns (eight narrow-band, two broad-band, one summary) and samples that would be represented by individual rows. In this manner,

one could determine not only if Anxious/Depressed is usually clinically elevated for students who receive solitary or combined anxiety and/or depression diagnoses, but that an Anxious/Depressed clinical elevation is infrequent in samples that are not characterized by either depression or anxiety. A review of this literature may also demonstrate which scales may be used to differentiate between conditions frequently addressed in the diagnostic process, as well as external correlates of these scales. (Apparently some of the best available published support for scale validity is confined to contrasted group analysis instead of continuous heterogeneous clinical samples and to the collection of nonindependent criteria from the same informant who completed the CBCL measure: see for example, Edelbrock & Costello, 1988; Weinstein, Noam, Grimes, Stone, & Schwab-Stone, 1990.)

4.13.2.4 Behavior Assessment System for Children

[American Guidance Service Inc., 4201 Woodland Road, Circle Pines, MN 55014-1796, USA]

The Behavior Assessment System for Children (BASC) differs from the three other systems presented in this chapter in that the entire system (Parent Rating Scales, PRS; Teacher Rating Scales, TRS; Self-report of Personality, SRP) was published at one time. (The BASC also includes a developmental history form and a classroom observation form, the Student Observation System). The BASC conveniently provides one integrated manual for all rating instruments (Reynolds & Kamphaus, 1992). Considering that these materials first became available in 1992 (with the advanced computer program first available in 1994), its application in the literature has been limited, although assessment applications continue to be developed. For example, the BASC Behavior Monitor for ADHD, brief parent and teacher forms intended for repeated use to demonstrate whether the behaviors of children already diagnosed as ADHD change as a result of treatment, became available in 1998.

In a manner similar to the Achenbach and the Conners' scales, the BASC ratings completed by parent, teacher, and youth are marked directly on self-scoring pamphlets or on one-page forms that allow the recording of responses for subsequent computer entry. Each of these forms is relatively brief (126–186 items) and can be completed in 10–30 minutes. PRS and TRS items are rated on a four-point frequency scale (never, sometimes, often, and almost

always), while SRP items are rated as either true or false. Raw scale scores are easily obtained from the self-scoring form, although selecting and transferring scale and composite standard scores in the form of linear *T* scores from 77 pages of conversion tables appears to require significant effort and concentration. Computer scoring and interpretation greatly facilitate this process.

For all three informants, scales are provided that measure clinical dimensions and adaptive dimensions (positive, desirable). All items are placed on only one substantive scale following an iterative process in which item content, item to scale correlations, normative-clinical endorsement differences and factor structure determined scale structure. In addition, these scales are combined on the basis of their intercorrelations into broad-band measures or scale composites, Externalizing Problems, Internalizing Problems, and Adaptive Skills. The relative diagnostic utility of scales vs. composites remains to be demonstrated. Children of school age are evaluated by the use of parent (PRS) and teacher (TRS) forms designated for ages 6–11 or 12–18. Scale item composition differs by both informant and age group, although substantial similarity is demonstrated across age groups within each informant-specific form. The Self-report of Personality (SRP) is similarly divided into forms for ages 8–11 and 12–18. In contrast to the PRS and TRS that emphasize across-informant similarities, the SRP has been designed to complement parent and teacher reports as a measure focused upon emotions and self-perceptions rather than as reports of overt behaviors.

BASC forms are also notable for their incorporation of validity scales to estimate the accuracy of obtained standard scores as well as provision of a brief Critical Items set. (The BASC manual does not, however, explain the rationale for item selection or provide relevant item performance statistics for these critical items.) Three sets of norms are provided: gender-specific (including representative proportions of students receiving special education services), combined gender, and clinical.

The use of a combined gender normative sample to generate standard scores is the exception to general assessment practice, as most measures of adjustment and personality compare each student to same-gender norms. It would seem that such a procedure would further minimize the significance of disruptive behaviors in girls and internalizing symptoms in boys, as well as present boys as relatively lacking in adaptive competence when compared to girls (see BASC manual, Figure 11.2). One could easily calculate the effect of any cutting point,

such as $T > 60$, on each BASC scale using both normative approaches within a sample of referred students. After students would be so identified, the accuracy of this classification could be established using information independent of BASC data.

In contrast to the BASC comprehensive normative sample, the sample used to generate clinical standard scores appears to both be limited in size (in some age/form samples less than 100 students) and restricted in composition. Yet test materials provide complete support for scoring results on clinical norms. Clinical norms are only routinely used for assessment measures that cannot be effectively applied within general population samples (e.g., a rating scale for severely disturbed hospitalized patients in which 99% of a regular population would obtain a summary score of 0). It is also difficult to define the parameters necessary to construct a "representative" clinical sample. Certainly TRS, PRS, and SRP samples of 30 or fewer children diagnosed as depressed or autistic would be inadequate for anything but initial exploration of scale or composite construct validity.

4.13.2.4.1 BASC Parent Rating Scales

PRS scale composites and associated narrow-band scales are Internalizing Problems (Anxiety, Depression, and Somatization), Externalization Problems (Hyperactivity, Aggression, and Conduct Problems), and Adaptive Skills (Adaptability (ages 6–11 only), Social Skills, and Leadership). Additional profiled scales include Atypicality, Withdrawal, and Attention Problems. The PRS also provides a Behavioral Symptoms Index, consisting of those clinical scales common to all age levels on both the PRS and TRS that load the highest on the first unrotated factor in a common-factor analysis. This combination of scales is presented in the manual as a measure of overall level of problem behavior. The BASC manual suggests that clinical scales should be considered of potential significance when $T > 60$, while adaptive scale scores should receive diagnostic attention when $T < 40$.

4.13.2.4.2 BASC Teacher Rating Scales

TRS scale composites and associated narrow-band scales include Internalizing Problems and Externalization Problems composites with the same narrow-band scales as for the PRS. The TRS Adaptive Skills composite includes the PRS scale components as well as a Study Skills scale that provides unique classroom observation (e.g., "Reads assigned chapters," "Studies

with other students"). Both TRS forms provide a School Problems Composite that incorporates scales Attention Problems and Learning Problems ("Does not complete tests," "Makes careless errors"). In this manner the TRS provides 21 or 22 items that are unique to the classroom.

Both the PRS and TRS incorporate a "fake bad" (F) index to assess the possibility that a teacher or parent rated a child in an inordinately negative fashion. The F index includes 16–20 items per form that represent either maladaptive behaviors to which the respondent answered "Almost always" or adaptive behaviors to which the respondent answered "Never." These extreme scores (neither item nor scale statistics are provided in the manual) may represent either problem exaggeration or the presence of severe behavioral disturbance.

4.13.2.4.3 BASC Self-Report of Personality (SRP)

SRP scales and scale composites vary from both PRS and TRS dimensions. Narrow-band scales Attitude to School, Attitude to Teachers, and (only for ages 12–18) Sensation Seeking form the School Maladjustment Composite. Scales Atypicality, Locus of Control, Social Stress, Anxiety, and (only for ages 12–18) Somatization form the Clinical Maladjustment Composite. A Personal Adjustment Composite is formed by scales Relations with Parents, Interpersonal Relations, Self-Esteem, and Self-Reliance. Two other unassigned narrow-band scales are included: Depression and Sense of Inadequacy. An additional measure, the Emotional Symptoms Index (ESI) is, according to the BASC manual, the SRP's most global indicator of serious emotional disturbance, particularly internalized disorders. The ESI is composed of two scales from the Clinical Maladjustment composite (Social Stress and Anxiety), two scales from the Personal Adjustment composite (Interpersonal Relations and Self-Esteem), and the two clinical scales that do not appear on any SRP composite (Depression and Sense of Inadequacy). Interpretive guidelines presented in the BASC manual generally suggest that scores of 60–69 T suggest an at-risk or mild/moderate range and scores above 69 T a "clinical range" interpretation for clinical scales. Because the adaptive scales are worded and scored in the positive direction, low elevations are given interpretive significance, at comparable ranges of 31–40 T and below 31 T .

The SRP also provides three scales to measure the presence of response sets that are likely to compromise scale validity. As in the PRS and TRS, the SRP includes an F-Index

composed of infrequent responses that may reflect symptom exaggeration (“Nobody likes me,” “Nothing about me is right”). In addition, a 14-item L-Index has been incorporated to measure problem denial (“I always go to bed on time,” “My parents are always right”). The third measure, V-Index, is not a scale but a small set of items that, when responded to in the True direction, are likely to reflect either lack of cooperation or inadequate comprehension (“I have never been in a car,” “Television does not really exist”).

4.13.2.4.4 *BASC Monitor for ADHD*

The BASC Monitor (Kamphaus & Reynolds, 1998) consists of brief parent and teacher forms that, along with the BASC Student Observation System and other variables defined by the clinician, provide a multimethod system designed to evaluate the effectiveness of treatments used with ADHD. The primary purpose is to demonstrate whether the behaviors of children already diagnosed as ADHD change as a result of treatment. These forms are intended for repeated use with the same child. Items are derived from 1992 PRS and TRS forms, other items collected during form development but not placed on a 1992 scale, as well as a few items written for this specific purpose that reflect *DSM-IV* diagnostic criteria. Norms were directly derived or extrapolated from BASC data and are organized into five age groups in order to be sensitive to age effects: 4–5, 6–7, 8–11, 12–14, and 15–18.

The Parent Monitor Ratings (PMR) consists of 46 items and the Teacher Monitor Ratings (TMR) consists of 47 items. All items are rated on the BASC four-choice frequency format of Never, Sometimes, Often, or Almost Always. Each form provides scales named Attention Problems, Hyperactivity, Internalizing Problems, and Adaptive Skills, and a listing of *DSM-IV* Items, some of which appear on other PRS and TRS scales, had been studied but not retained, or were written specifically to reflect *DSM* criteria. Item content was selected to be appropriate across the 4–18 years span, and some variation between PMR and TMR was allowed to reflect the unique opportunities for observation associated with each rater (e.g., TMR: “Bothers other children when working,” PMR: “Fiddles with things while at meals”).

4.13.2.4.5 *Commentary*

It is difficult for a newly published set of instruments to compete with those with a 20- to 30-year history of application. Certainly the BASC cannot draw upon an extensive research

literature to demonstrate scale validity and to suggest appropriate clinical and research applications. The BASC family of instruments are attractive for several reasons, however, especially if all three ratings (parent-, teacher-, and self-report) are obtained for a student. Scale and composite distributions appear to be less skewed than many Achenbach and Conners' measures. Cluster analysis of the TRS normative sample, for example, resulted in seven clusters. In these clusters mean clinical scale *T* scores ranged from a low 40s to a high at or above 70, while mean adaptive scale *T* scores ranged from a low of 33 to a high of 60 (Kamphaus, Huberty, DiStefano, & Petoskey, 1997). The second reason, SRP *T* scores easily are obtained below 50*T*, in contrast to the YSR, and substantial item content is usually represented in *T* scores within the “clinical range.” Another reason to consider these scales is that application of the BASC in school psychology settings in which ease of collection of both teacher and parent ratings is greater than in other settings would take advantage of the complementary nature of the SRP when the PRS and TRS are routinely collected. The final reason to consider the BASC is that the SRP scales appear especially attractive in the dimensions they generate, if parent and teacher ratings are available to document comorbid disruptive behavior phenomena. For example, it would be clinically meaningful in an adolescent described by parent and teacher as aggressive and conduct disordered if he would describe himself as seeing others as responsible for his problems (Locus of Control) and as focused on the experiences of thrill-seeking and excitement (Sensation Seeking).

As with any first manual, however, the psychologist will be left with many questions regarding application of BASC scales to the individual student. First, it is unclear from what source interpretive guidelines are derived, unless from scale item content. For example, elevation of SRP Sensation Seeking is presented to be associated with potential for alcohol and drug use or experimentation, while elevation of SRP Social Stress in young children with scores higher than 70*T* suggests children who “may turn inward in an unsuccessful attempt to cope with these tensions” (*BASC manual*, pp. 60–61). Perhaps these conclusions are not derived from scale correlates, but from an extrapolation of the phenomena associated with the clinical dimension measured by a given scale.

Although validity scales are provided, especially for SRP, the psychometric details of their construction and their performance in data presented in the manual are not provided. Does the L-Index correlate negatively with SRP clinical scales and positively with SRP adaptive

scales? Were the clinical samples presented in the manual first screened by these scales to exclude potentially invalid protocols? The evidence of scale validity primarily takes the form of factor structure and correlations with other published measures. Readers of the manual will not find a series of case studies, nor are tabled and profiled mean scale values for a variety of referred samples presented in terms of proportion that exceed some value presented as clinically meaningful. In this regard, it is most interesting to contrast manual tables 12.28, 13.25, and 14.22 in that not one SRP clinical scale obtains a mean of at least 60*T*, while only one adaptive scale score obtains a mean of 40*T* or less. Similar normal limits results were obtained for a relatively small sample of children in state custody due to abuse or neglect, although the L-Index was apparently not scored (Dalton, 1996).

A brief review of published studies using the PRS and TRS, in contrast, provides some encouraging evidence of scale validity. BASC parent and teacher scales demonstrated convergent validity with CBCL and TRF scales (Vaughn, Black, Hall, Hynd, & Riccio, 1995). PRS scales were found to demonstrate convergent validity in nonreferred kindergarten students with scales of the Social Skills Rating System for Children (Flanagan, Alfonso, Primavera, & Povall, 1996) and demonstrated similar validity in elementary school students at risk for conduct disorder with scales of the CBCL (Doyle, Ostrander, Skare, Crosby, & August, 1997). In a study of cognitive problem-solving, increasing PRF Hyperactivity Scale elevation was associated with more problem solutions with aggressive or noncompliant content as well as fewer reasons or consequences justifying the selection of the "best solution" (Bloomquist, August, Cohen, Doyle, & Everhart, 1997). Four TRS scales (Aggression, Conduct Problems, Depression, and Social Skills) significantly classified ADHD and non-disabled students, as well as ADHD students with and without a comorbid condition (Lett & Kamphaus, 1997). TRS Learning Problems correlated in referred elementary students with the Freedom from Distractibility factor, although no relation was obtained for the Attention Problems or Hyperactivity scales (Lowman, Schwanz, & Kamphaus, 1996). In addition, TRS Externalizing Problems correlated significantly with a peer rating dimension "aggressive-disruptive," while TRS Internalizing Problems correlated significantly with a peer rating dimension "sensitive-isolated" (Realmuto, August, Sieler, & Pessoa-Brandao, 1997). In a longitudinal study, teacher/parent composite measures of Hyperactivity/Attention Pro-

blems and Aggression/Conduct Problems were strongly intercorrelated and negatively related to adaptive functioning 3.5 years later (August, MacDonald, Realmuto, & Skare, 1996).

4.13.2.5 Personality Inventory for Children, 2nd Edition (PIC-2); Personality Inventory for Youth (PIY); Student Behavior Survey (SBS)

[Western Psychological Services, 12031 Wilshire Boulevard, Los Angeles, CA 90025-1251, USA]

Forty years ago two University of Minnesota psychologists began the development of a new inventory approach to the evaluation of children and adolescents. They assembled a 600-item administration booklet and named it the "Personality Inventory for Children. For use with children from six through adolescence." The directions stated that each item was to be answered as either "true" or "false" by the child's mother in order to describe both the child and family relationships. Professors Wirt and Broen accumulated administration booklet descriptive statements following an outline. To ensure comprehensive coverage of child behavior and adjustment, 50 items were sorted into 11 separate content areas: aggression, anxiety, asocial behavior, excitement, family relations, intellectual development, physical development, reality distortion, social skills, somatic concern, and withdrawal. To these 550 potential scale items, 50 items were added in an effort to strengthen or clarify the meaning of certain areas of concern. Following many of the general procedures employed in development of the Minnesota Multiphasic Personality Inventory, PIC scales were identified from this item pool over a span of 20 years.

The initial 1977 published profile included three validity scales, a general screening scale, and 12 measures of child ability and adjustment and family function developed through either empirical item selection techniques or through iterative content valid procedures. In 1981 the administration booklet was revised and items were sorted into one of four parts. Completion of part I (items 1-131) allowed the scoring of four broad-band factor-derived scales (Lachar, Gdowski, & Snyder, 1982). Completion of parts I and II (items 1-280) generated the entire clinical profile with "shortened scales" (Lachar, 1982), while parts I-III (420 items) allowed the scoring of the original length scales. The final 180 items of this booklet were eventually dropped from the administration booklet because they did not appear on any of the standard full-length profile scales.

From the beginning, the task of PIC scale and profile interpretation attributed relatively little importance to item content (except for the construction of a Critical Items list), but instead established external correlates and interpretive guidelines for individual profile scales (Lachar & Gdowski, 1979) and replicated profile patterns (Gdowski, Lachar, & Kline, 1985; Kline, Lachar, & Gdowski, 1987; LaCombe et al., 1991). A diagnostic procedure in which similarity coefficients are calculated between the individual PIC profile to be interpreted and the mean profiles of students receiving specific special education services has also been incorporated into profile scoring and interpretation software (Kline, Lachar, Grubber, & Boersma, 1994). Special effort has focused on determining that PIC scale validity is not restricted by a child's age, gender, or ethnicity status (Kline & Lachar, 1992; Kline et al., 1985).

The development efforts for the PIC-2 started in 1989 with the rewriting of the first 280 items of the PIC booklet into a self-report format for the PIY (Lachar & Gruber, 1993, 1995). Many dimensions were easily translated from the "my child" to the "I" format, although this process was either difficult or impossible for some dimensions. For example, accurate self-report of early developmental phenomena and severe developmental delay would be unlikely or inappropriate for students who can complete a comprehensive self-report inventory ("My child was difficult to toilet train," "My child can comb his/her own hair"). Such PIC items were replaced with other self-report statements likely to be used in validity scales or to supplement coverage of dimensions such as depression and defective reality testing (see PIY technical manual for an elaboration of this process). The PIY was developed with a subscale-within-scale structure that facilitates profile interpretation and has also incorporated a set of validity scales constructed from the responses of children and adolescents.

Development of the PIY facilitated concurrent critical review of the structure and content of PIC scales and the PIC profile. This review motivated substantial data collection using a PIC research edition administration booklet that allowed both the scoring of the PIC and the collection of data on revised and new inventory items. Data collected from application of this booklet have been subjected to considerable statistical analysis through which the PIC (now PIC-2) and PIY have achieved a great deal of similarity to facilitate comparison. The PIC validity scales, already a significant component of profile interpretation, have been improved in the PIC-2 and one consistent clinical scale and subscale construction methodology has been

applied to the assessment of all revised clinical dimensions. In addition, a national normative sample has been collected concurrently with that for the third diagnostic component, a new teacher rating form, the Student Behavior Survey (SBS). Although these revisions have increased the focus on content validity in this family of parent, teacher, and self-report questionnaires, the 20-year tradition of deriving interpretive guidelines from the empirical relations between test measures and independent phenomena of adjustment and behaviour continues in the PIC-2 and SBS.

4.13.2.5.1 Personality Inventory for Children, 2nd Edition (PIC-2); Personality Inventory for Youth (PIY)

The PIC-2 and PIY profiles graph the linear *T* scores of three parallel validity scales and nine substantive clinical scales. PIC-2 norms are provided for students in grades K through 12 (a preschool version is under development), while PIY norms are available for students in grades 4 through 12. Clinical scales were constructed using an iterative process in which initial item placement based upon previous PIC scale placement and/or substantive item content was then empirically supported within item-to-scale correlation matrices based upon large clinical samples. Items retained in each final PIC-2 and PIY clinical scale demonstrated both statistical significance and the most substantial item-to-scale correlation among the nine potential scale placements. All other items, unless placed on an inventory validity scale or a critical item list, were dropped from the final version of each inventory. This process, in contrast to item selection through contrasted groups, placed an emphasis on item content and scale homogeneity.

The majority of PIY clinical scale items (86–100%, mean = 95%) appear on the identically named PIC-2 clinical scale, facilitating comparison of parent- and youth-report. The items of the PIY clinical scales are placed on only one scale and one subscale. Of the 264 items that comprise the nine PIC-2 clinical scales, only 16 (6.1%) appear on two scales. Each clinical scale is also divided into two or three nonoverlapping subscales. Guided by a series of item factor analyses, these subscales consist of item subsets that are more homogeneous than their associated total scales. Clinical and validity scale content remains the same across child age or gender, allowing expected age- and gender-related variance to appear. (For example, greater elevation on the Impulsivity and Distractibility scale will be found for boys in comparison to girls and for

younger children in comparison to older children.) PIY clinical scales become interpretable at 60*T*, and subscales usually at 65*T*. (Analyses of PIC-2 scale standard scores suggest comparable guidelines.) Table 3 presents the scale/subscale structure, representative items, and internal consistency estimates of the PIC-2 parent-report and PIY self-report clinical scales.

Table 3 demonstrates that PIC-2 and PIY item content varies from the description of commonly occurring difficulties to statements suggestive of severe psychopathology. It should be noted that many PIC-2 and PIY subscales incorporate more descriptors than the full scales found in the other three families of objective questionnaires.

The PIY and PIC-2 both incorporate a screening or short assessment procedure. The first 80 items of the PIY comprise a 32-item screening scale chosen to provide an optimal identification of those regular education students who—when administered the full PIY—produced clinically significant results. These items also include three “scan items” for each scale. Scan items were selected in such a manner so that individuals who endorse two or more of each set of three items would be those with a high probability of scoring > 59*T* on the corresponding clinical scale. Shortened versions of three validity scales can also be derived from these items.

The PIC-2 provides a short form to measure change in clinical status associated with therapeutic intervention. Although PIC scales may have demonstrated such sensitivity to change (see therapeutic case example in Lachar & Kline, 1994), a brief form tailored specifically for this purpose was constructed. Selected items were (i) written in the present tense, (ii) frequently endorsed in the context of clinical assessment, and (iii) described clinical phenomena often the focus of short-term intervention. Using these guidelines, the 12 most favourable items from each of eight PIC-2 clinical scales were selected. (Cognitive Impairment was excluded due to historical components, lack of appropriate therapeutic focus, and the global/stable nature of most descriptions on this dimension.) These 96 inventory statements have been placed at the beginning of the 275-item PIC-2 booklet to serve as both a short form and a method of efficient re-evaluation of a child following short-term intervention. These 96 items are also available in a self-scoring format. It is intended for these scale scores to be graphed on the same profile at baseline and at appropriate interim and post-treatment intervals to demonstrate both dimensions of change and stability.

The initial concurrent validity of these shortened scales was established through correlation of these scale scores with clinician ratings,

teacher descriptions, and self-report descriptions. These data were drawn from an ongoing clinical project in which the PIY, SBS, clinician ratings, diagnoses, and individually administered psychometric results are being collected to some degree with over 1000 PIC-2 assessments. Even though an exploratory analysis, some care was taken in the selection of these correlates. Each obtained scale descriptor was placed on only the one scale with which it received the largest correlation, all being at least significant at $p < 0.01$. Table 4 summarizes the number of ratings identified in this manner from each source (clinician, teacher, and student) and provides three examples from each rating source for each of these eight scales. Correlations between these shortened scales and their full-length versions are also presented.

Table 4 demonstrates that these brief scales correlated substantially with full-length versions and obtained independent correlates from nonparent observers that matched expressed scale content and diagnostic intent. Clinicians provided the greatest support for ADH, DLQ, and DIS, focusing on problems of disruptive and noncompliant behavior and intense and dysphoric affect that often form the basis of clinical referral. These analyses demonstrate that ADH, as previously demonstrated for the PIC Hyperactivity scale (Lachar & Gdowski, 1979), assesses those behaviors most related to problems in classroom adjustment. In addition, observations obtained directly from the student under study provide those internal and subjective judgments that demonstrate the clinical value of PIC-2 dimensions which do not receive robust correlates from clinicians or teachers.

PIY and PIC-2 profiles incorporate three validity scales. The Inconsistency scale (INC) evaluates the likelihood that responses to items are random or reflect in some manner inadequate comprehension of inventory statements or compliance with test instructions. The Defensiveness scale (DEF) identifies profiles likely to demonstrate the effect of minimization or denial of current problems. The third validity scale, Dissimulation (FB), identifies profiles that may result from either exaggeration of current problems or a malingered pattern of atypical or infrequent symptoms.

PIY and PIC-2 Inconsistency scales measure semantic inconsistency through the classification of response to 35 pairs of highly correlated items drawn from all nine clinical scales (for example, “I have many friends/I have very few friends”; “My child has a lot of talent/My child has no special talents”). For each item pair, two response combinations are consistent and two are inconsistent. Each inconsistent pair

Table 3 PIC-2 and PIY Clinical Scale and Subscale Composition.

Cognitive Impairment	PIC-2: 39 items, alpha = .881 PIY: 20 items, alpha = .740 Items in common: 20 (only 1 on PIC/2 COG3)
PIC-2 Subscales:	
COG1/Inadequate Abilities (13 items, alpha = .792)	My child is rather absent-minded.
COG2/Poor Achievement (13 items, alpha = .785)	Reading has been a problem for my child.
COG3/Developmental Delay (13 items, alpha = .799)	My child could eat with a fork before age four years.
PIY Subscales:	
COG1/Poor Achievement and Memory (8 items, alpha = .649)	School has been easy for me.
COG2/Inadequate Abilities (8 items, alpha = .673)	Other people think that I am talented.
COG3/Learning Problems (4 items, alpha = .441)	I have been held back a year in school.
Impulsivity and Distractibility	PIC-2: 27 items, alpha = .921 PIY: 17 items, alpha = .773 Items in common: 17
PIC-2 Subscales:	
ADH1/Disruptive Behavior (21 items, alpha = .917)	My child jumps from one activity to another.
ADH2/Fearlessness (6 items, alpha = .693)	My child will do anything on a dare.
PIY Subscales:	
ADH1/Brashness (4 items, alpha = .535)	I like to show off.
ADH2/Distractibility/Overactivity (8 items, alpha = .613)	I cannot keep my attention on anything.
ADH3/Impulsivity (5 items, alpha = .543)	I often act without thinking.
Delinquency	PIC-2: 47 items, alpha = .957 PIY: 42 items, alpha = .923 Items in common: 39
PIC-2 Subscales:	
DLQ1/Antisocial Behavior (13 items, alpha = .881)	My child has been in trouble with the police.
DLQ2/Dyscontrol (17 items, alpha = .921)	When my child gets mad, watch out!
DLQ3/Noncompliance (17 items, alpha = .927)	My child often disobeys me.
PIY Subscales:	
DLQ1/Antisocial Behavior (15 items, alpha = .831)	I have run away from home.
DLQ2/Dyscontrol (16 items, alpha = .839)	What people say often makes me angry.
DLQ3/Noncompliance (11 items, alpha = .828)	I tend to see how much I can get away with.
Family Dysfunction	PIC-2: 25 items, alpha = .882 PIY: 29 items, alpha = .869 Items in common: 25
PIC-2 Subscales:	
FAM1/Conflict Among Members (15 items, alpha = .847)	Our family argues a lot at dinner time.
FAM2/Parent Maladjustment (10 items, alpha = .781)	One of the child's parents drinks too much alcohol

Table 3 (continued).

PIY Subscales:

FAM1/Parent–Child Conflict (9 items, $\alpha = .817$)

I am unhappy about my home life.

FAM2/Parent Maladjustment (13 items, $\alpha = .738$)

My parents often argue.

FAM3/Marital Discord (7 items, $\alpha = .701$)

My parents are now divorced or living apart.

Reality Distortion PIC-2: 29 items, $\alpha = .898$ PIY: 22 items, $\alpha = .831$

Items in common: 20

PIC-2 Subscales:

RLT1/Developmental Deviation (14 items, $\alpha = .842$)

My child does not understand other people.

RLT2/Hallucinations and Delusions (15 items, $\alpha = .817$)

My child sometimes sees things that are not there.

PIY Subscales:

RLT1/Feelings of Alienation (11 items, $\alpha = .768$)

I am different from most kids.

RLT2/Hallucinations and Delusions (11 items, $\alpha = .709$)

People secretly control my thoughts.

Somatic Concern PIC-2: 28 items, $\alpha = .829$ PIY: 27 items, $\alpha = .853$

Items in common: 24

PIC-2 Subscales:

SOM1/Psychosomatic Preoccupation (17 items, $\alpha = .793$)

My child seems tired most of the time.

SOM2/Muscular Tension and Anxiety (11 items, $\alpha = .675$)

Recently my child has complained of chest pains.

PIY Subscales:

SOM1/Psychosomatic Syndrome (9 items, $\alpha = .730$)

I often have headaches.

SOM2/Muscular Tension and Anxiety (10 items, $\alpha = .740$)

At times I have trouble breathing.

SOM3/Preoccupation with Disease (8 items, $\alpha = .596$)

I am worried about disease.

Psychological Discomfort PIC-2: 39 items, $\alpha = .909$ PIY: 32 items, $\alpha = .864$

Items in common: 31

PIC-2 Subscales:

DIS1/Fear and Worry (13 items, $\alpha = .728$)

My child is often afraid of little things.

DIS2/Depression (18 items, $\alpha = .881$)

My child tends to feel sorry for himself/herself.

DIS3/Sleep Disturbance/Preoccupation with Death (8 items, $\alpha = .770$)

My child's sleep is calm and restful.

PIY Subscales:

DIS1/Fear and Worry (15 items, $\alpha = .779$)

I worry about things that adults worry about.

DIS2/Depression (11 items, $\alpha = .730$)

The future looks good to me.

DIS3/Sleep Disturbance (6 items, $\alpha = .703$)

I often get up at night.

Social Withdrawal PIC-2: 19 items, $\alpha = .796$ PIY: 18 items, $\alpha = .800$

Items in common: 18

Table 3 (continued).

PIC-2 Subscales:
WDL1/Social Introversion (11 items, $\alpha = .781$)
My child worries about talking to others.
WDL2/Isolation (8 items, $\alpha = .656$)
My child often stays in his/her room for hours.
PIY Subscales:
WDL1/Social Introversion (10 items, $\alpha = .781$)
Shyness is my biggest problem.
WDL2/Isolation (8 items, $\alpha = .591$)
I keep my thoughts to myself.
Social Skill Deficits PIC-2: 28 items, $\alpha = .915$
PIY: 24 items, $\alpha = .855$
Items in common: 23
PIC-2 Subscales:
SSK1/Limited Peer Status (13 items, $\alpha = .853$)
My child often brings friends home.
SSK2/Conflict with Peers (15 items, $\alpha = .888$)
Other children often get mad at my child.
PIY Subscales:
SSK1/Limited Peer Status (13 items, $\alpha = .788$)
I am sure of myself in a group.
SSK2/Conflict with Peers (11 items, $\alpha = .804$)
Other kids make fun of my ideas.

Note: coefficient alphas based upon 1178 PIY and 901 PIC-2 protocols obtained in the context of clinical evaluation.

identified in a given protocol contributes one point to the INC raw score. Application of a cutting raw score of > 12 resulted in correct identification of 90–95% of clinical protocols and 92–96% of a random sample. The DEF is an expanded version of the PIC Lie scale. DEF items represent denials of common problems (“Sometimes I put off doing a chore. False”; “My child almost never argues. True”) and attributions of improbable positive adjustment (“My child always does his/her homework on time. True”; “I am almost always on time and remember what I am supposed to do. True”). DEF elevations above 59T, even in hospitalized psychiatric patients, result in profiles that either minimize current problems or consistently deny the presence of most or all problems in adjustment.

The third validity scale, Dissimulation (FB), was empirically constructed through item analysis comparing clinical protocols and two sets of protocols completed by nonreferred regular education students or their mothers. First the PIY or PIC-2 was completed with directions to provide an accurate description, then a second questionnaire was completed as if the student was in need of mental health counseling or hospitalization. Selected items in the scored direction were very infrequent in the accurate normal (PIY, 11%; PIC-2, 4%) and clinical (PIY, 18%; PIC-2, 15%) condition,

while characteristic (PIY, 83%; PIC-2, 55%) of the “fake bad” or dissimulated condition. The pattern of these three scales readily identifies profiles in which caution must be applied to their interpretation. Table 5 presents PIC-2 and PIY profile scores from three hospitalized 12-year-old adolescents.

The evaluation of Case A resulted in PIC-2 and PIY validity scale results which did not suggest that the accuracy of these profiles had been compromised in any systematic manner. It is important to observe that both profiles present clinically elevated scale scores. First consider similarities in scale and subscale pattern and scale and subscale clinical elevations (e.g., apparent disagreement for COG3 represents the reality that PIC-2 COG3 consists of parent report of developmental delay, while PIY COG3 represents learning problems—a dimension much more similar to PIC-2 COG2). Case A is a 12-year-old in his fifth psychiatric hospitalization who had been placed in a self-contained special education classroom for behavioral maladjustment. He carries diagnoses of ADHD combined type, Oppositional Defiant Disorder, and Conduct Disorder. His behavior at home and in the hospital demonstrated serious behavioral dyscontrol. He had been noncompliant in taking psychotropic medication to improve his emotional and behavioral control. He had threatened to kill himself (see

Table 4 Correlates of PIC-2 Brief Clinical Scales.

Impulsivity and Distractibility (ADH12: .96 ^a)
Clinician Ratings: total = 14
Hyperactive, overactive
Restless
Frequently frustrated
Teacher Ratings: total = 34
Disobeys class or school rules
Impulsive, acts without thinking
Misbehaves unless closely supervised
Student Ratings: total = 9
Brags about being sent to the principal
School has sent notes home about bad behavior
Likes to show off
Delinquency (DLQ12: .93)
Clinician Ratings: total = 44
Argues, Oppositional
Disobedient to parents
Poorly modulated anger
Teacher Ratings: total = 11
Does not demonstrate polite behavior
Blames others for his/her problems
Becomes upset for little or no reason
Student Ratings: total = 24
Gives parents a lot of trouble
Sometimes swear at parents
Ran away from home
Family Dysfunction (FAM12: .93)
Clinician Ratings: total = 6
Conflict between parents
Parent divorce/separation
Emotionally abused
Teacher Ratings: total = 6
Uses alcohol or drugs
Strikes or pushes school personnel
Preoccupied with sex
Student Ratings: total = 18
Family does not enjoy being with each other more than other families
A lot of tension in the home
Parents often argue
Reality Distortion (RLT12: .95)
Clinician Ratings: total = 3
Auditory hallucinations
Inappropriate emotion, affect
Delusions, paranoia
Teacher Ratings: total = 0
Student Ratings: total = 13
Don't get along with others most of the time
Need a lot of help from others
Do strange or unusual things
Somatic Concern (SOM12: .92)
Clinician Ratings: total = 6
Excessive sleeping
Sexually abused
Continually tired, listless

Table 4 (continued).

Teacher Ratings: total = 0
Student Ratings: total = 33
Often get very tired
At times have trouble breathing
Don't have as much energy as most other kids
Psychological Discomfort (DIS12: .93)
Clinician Ratings: total = 22
Depressed, sad, unhappy
Moodiness
Inadequate self-esteem
Teacher Ratings: total = 4
Appears sad or unhappy
Doesn't take successes and failures in stride
Worries about little things
Student Ratings: total = 25
Shy with kids my own age
Not often in a good mood
Often afraid of little things
Social Withdrawal (WDL12: .96)
Clinician Ratings: total = 4
Withdrawn
Shy
Uncommunicative, seldom talks
Teacher Ratings: total = 2
Pessimistic about the future
Overcritical of himself/herself
Student Ratings: total = 12
Shyness is my biggest problem
Hardly ever talk
Stay in the house for days at a time
Social Skill Deficits (SSK12: .96)
Clinician Ratings: total = 5
Teased by peers
Isolated, few or no friends
Attends self-contained special education class
Teacher Ratings: total = 17
Does not seem to have fun
Avoids social interaction in class
Unaware of the feelings of others
Student Ratings: total = 28
Often rejected by other kids
Have very few friends
Not very popular with other kids

^a Indicates correlation between total and 12-item version of PIC-2 clinical scale in a sample of 905 referred students. Note text for explanation of correlate selection procedure.

DIS3 scores) and had assaulted his mother who he said did not want him at home (see PIY FAM1 and FAM2). This patient attempted to escape from the hospital and required multiple time-outs and seclusions to control rages, threats, and aggressive and inappropriate behavior (see DLQ2). Off medication, A demonstrated an attention span of less than

15 minutes (see ADH values). He had a history of impulsive and disruptive behavior, fighting with peers, noncompliance with adults, verbal and physical aggression, and running away from home (elevated DLQ values for PIY and PIC-2).

Case B, in contrast to Case A, documents considerable parent/child disagreement. The

Table 5 The influence of respondent defensiveness on PIY/PIC-2 profile pairs.

<i>Scale/subscale</i> <i>NORM</i>	<i>Case A</i>		<i>Case B</i>		<i>Case C</i>	
	<i>PIC-2</i>	<i>PIY</i>	<i>PIC-2</i>	<i>PIY</i>	<i>PIC-2</i>	<i>PIY</i>
Inconsistency	53	49	67	57	50	68
Dissimulation	60	72	81	48	47	72
Defensiveness	30	39	30	64	65	50
Cognitive Impairment	61	63	67	57	43	47
COG1	56	57	68	52	49	42
COG2	67	50	70	45	41	45
COG3	53	85	48	85	43	67
Impulsivity and Distractibility	79	69	81	39	44	64
ADH1	75	77	83	49	46	53
ADH2	81	62	64	37	41	62
ADH3		57		41		66
Delinquency	83	75	86	41	42	49
DLQ1	71	75	63	43	46	49
DLQ2	82	71	98	41	43	56
DLQ3	79	67	76	43	42	42
Family Dysfunction	60	69	58	57	67	49
FAM1	58	69	58	52	62	47
FAM2	60	74	54	57	72	49
FAM3		53		58		53
Reality Distortion	67	71	73	48	50	79
RLT1	69	64	83	53	55	74
RLT2	62	75	56	41	44	79
Somatic Concern	82	75	74	49	41	65
SOM1	82	66	76	38	42	59
SOM2	72	71	65	53	41	65
SOM3		73		59		65
Psychological Discomfort	90	59	97	63	53	68
DIS1	70	64	81	64	51	66
DIS2	88	38	92	52	49	58
DIS3	83	70	83	63	63	68
Social Withdrawal	47	83	85	41	46	59
WDL1	46	77	72	46	49	53
WDL2	50	78	88	38	42	65
Social Skill Deficits	57	43	86	56	49	57
SSK1	43	40	68	53	50	56
SSK2	74	50	97	59	48	55

Note: Table 3 provides the names and examples of PIC-2 and PIY subscales.

only clear agreement as to problem area is in academic achievement (PIC-2 COG2, PIY COG3). History and psychometric assessment document retention in grade, special class placement, and achievement substantially below assessed ability. Clearly the PIC-2 most accurately describes a 12-year-old male who demonstrates multiple handicaps. The elevation of the PIY DEF scale ($T = 64$) is the most likely explanation for a PIY profile that is essentially within normal limits. Indeed, review of the

medical record details this patient's repeated denial and minimizing of his problems during this hospitalization in an attempt to facilitate his early discharge from treatment. This child's psychiatric history was secondary to a traumatic motor vehicle accident. He felt lonely and scared, he frequently cried, sobbed, shook, avoided others, and was preoccupied with excessive worries (DIS, WDL). He externalized his problems (DLQ) and had difficulties with peers and had little insight into his role in

these conflicts (SSK2). This pattern of youth defensiveness is fairly common in inpatient settings.

Case C is quite unusual in that the mother-completed PIC-2 is essentially within normal limits, with the exception of FAM1, FAM2, and DIS3 obtaining some elevation. This normal-limits profile is quite unusual in light of the fact that the mother referred her 12-year-old daughter for this hospitalization. C presented with suicidal ideation, low self-esteem, depression, crying spells, poor appetite, and associated weight loss (DIS, WDL2). She actively demonstrated somatic concern and somatic symptoms in response to conflict during this hospitalization (SOM) and said that she would not talk about her problems with her mother, because she was afraid that such discussion would distress her mother who was under psychiatric care (PIC-2 FAM2). Clinicians were sufficiently concerned with C's internalizing problems to assign discharge diagnoses of Generalized Anxiety Disorder and Depressive Disorder NOS and to prescribe antidepressant medication. Why was C's mother more defensive in describing her daughter's problems than C was herself? It was clearly documented in the medical record that C's mother was concerned that, because of her and C's psychiatric problems, she would be seen as an inadequate mother and consequently would lose custody of her child to another adult family member.

4.13.2.5.2 Student Behavior Survey

The Student Behavior Survey (SBS) differs from the other teacher rating scales described in this chapter in that no particular effort was made to develop the SBS into a parallel teacher version of the PIC-2 and PIY. Indeed, the development of SBS items consisted of several iterations in which the authors focused on content appropriate to teacher observation. SBS items and their rating options appear on a self-scoring form. These items are sorted into content meaningful dimensions under 11 scale headings rather than presented in random order. The resulting items demonstrate a school focus in that 58 of 102 items specifically refer to in-class or in-school behaviors and judgments that can only be made by school staff (Wingenfeld, Lachar, Gruber, & Kline, 1998). The SBS consists of 102 items profiled on to 14 scales that assess student academic status and work habits, social skills, parental participation in the educational process, and problems such as aggressive or atypical behavior and emotional stress. Norms that generate linear *T*

scores are gender-specific and divided into two age groups: 5–11 and 12–18 years.

The SBS consists of three sections (Lachar et al., 1999). In the first section the teacher circles one rating out of five (Deficient, Below Average, Average, Above Average, Superior) for eight areas of achievement such as reading comprehension and mathematics that are summed for an estimate of current academic performance. The remaining SBS items are rated on a four-point frequency scale: Never, Seldom, Sometimes, and Usually. The next two scales consist of positively worded descriptions of student adaptive behaviors. Academic Habits (13 items: "Completes class assignments"; "Follows the teacher's directions") and Social Skills (eight items: "Helps other students"; "Participates in class activities") scales document positive behaviors. Ratings of parents are very school-specific in that the teacher is asked to judge the degree to which parents support the student's educational program (six items: "Parent(s) encourage achievement"; "Parent(s) meet with school staff when asked").

The seven Problems in Adjustment scales consist of negatively worded items organized into Health Concerns (six items: "Complains of headaches"; "Talks about being sick"); Emotional Distress (15 items: "Appears sad or unhappy"; "Worries about little things"); Unusual Behavior (seven items: "Seems lost or disoriented"; "Says strange or bizarre things"); Social Problems (12 items: "Criticized by other students"; "Avoids social interaction in class"); Verbal Aggression (seven items: "Insults other students"; "Teases or taunts other students"); Physical Aggression (five items: "Destroys property when angry"; "Starts fights with other students"); and Behavior Problems (15 items: "Disrupts class by misbehaving"; "Talks excessively"). In initial samples of 1173 regular education students and 601 students referred for evaluation or receiving special education services in grades K through 12, 99 of 102 items significantly ($p < 0.001$) separated these two samples. All items demonstrated that they had been placed on the scale with which each obtained the largest correlation. Scale scores of regular education and referred students obtained meaningful three-factor solutions.

Additional effort (Pisecco et al., 1998) resulted in three additional, 16-item nonoverlapping scales consisting of SBS items first nominated as fitting *DSM-IV* diagnoses of ADHD (combined type), Oppositional Defiant Disorder, and Conduct Disorder. Item-to-scale correlations and a three-factor solution of these 48 SBS items empirically supported the placement of these scale items.

4.13.2.5.3 Commentary

The revision of the PIC, the addition of a teacher rating scale, and the collection of a national representative normative sample for each have gone a long way to respond to the concerns raised by reviewers regarding the PIC's age (Kamphaus & Frick, 1996; Knoff, 1989; Merrell, 1994). The evaluations of the SBS and PIC-2 manuals and the review of their subsequent demonstrated ability to evaluate emotional adjustment at baseline and to quantify response to intervention will appear well into the twenty-first century.

The emphasis on evaluating response accuracy using validity scales and the empirical determination of interpretive guidelines continues to characterize these measures. Many psychologists unconvinced of the importance of these concepts may not value their contributions to assessment. Although the PIC has been reduced from 420 to 275 items, into which a set of subscales and a brief form have been incorporated, some clinicians may still judge the length of these questionnaires to be problematic. Although this chapter's author is obviously biased against such a position, it is certain that the breadth and depth of a measure's content establishes the potential boundaries for its utility. Even the 270 items of the PIY are easily completed in less than 45 minutes by children as young as students in the fourth grade. Efficiency has been improved by rejecting any item not actively used in the interpretive process as well as providing computer software for scoring and interpretation.

As either "new" or "improved" measures, the approximately 400 PIC-relevant publications (bibliography available from this author) at best now suggest diagnostic potential for these forms on the dimensions that retain the greatest similarity from original to revised formats. A great deal of effort will be necessary to establish the diagnostic utility of the new and revised forms to achieve the standards of the original inventory (Lachar & Kline, 1994). Such efforts have begun. For example, matched samples of adolescents with discharge diagnoses of either Conduct Disorder or Major Depression were correctly classified by PIY subscales in 83% of cases (Lachar, Harper, Green, Morgan, & Wheeler, 1996).

4.13.3 CONCLUSIONS

Clinician interest in applying objective questionnaires in the evaluation of children and adolescents continues to grow. These measures are efficiently obtained, easily scored, and at least appear to be relatively easy to interpret

(although a sophisticated knowledge of both developmental psychopathology and psychometrics is clearly a prerequisite). The routine application of such measures in clinics, hospitals, and private practices will continue to be a growing trend. This chapter documents that a number of quality measures are currently available to meet this need. Clinicians should review each set of measures to select the specific family of instruments best suited for them to be applied as a routine procedure, or the specific instruments to be applied to a specific clinical activity.

Routine collection of information that is both multidimensional and multisource will influence future research and clinical practice. Efforts to document the effect of comorbid conditions and to establish interpretive guidelines for individual profiles and sets of profiles will continue. Interpretive handbooks, similar to those available for popular adult personality inventories, will be developed to support training and clinical applications. Projects will identify research designs in which the meaning of scale values and profile patterns can be established through the analysis of independent criteria. Given the infrequency of studies designed to evaluate tests, as compared to studies that use tests to evaluate clinical phenomena, efforts will be made to identify and collect substantial data that will allow such analyses. Indeed, the concurrent application of parent, teacher, youth, and clinician forms, as demonstrated in Table 4, will allow the establishment of external correlates and the determination of scale standard score interpretive ranges.

Studies will routinely investigate the relative efficacy of informant sources (parent, teacher, and youth) to support a given diagnostic decision, as well as compare the instruments across these four families. Some early efforts include Pearson and Lachar's (1994) comparison of CBCL and PIC scales in their ability to predict the standard scores of the Vineland Adaptive Behavior Scale, and the comparison of BASC and CBCL forms in identifying students with ADHD and in differentiating between ADHD types (Ostrander, Weinfurt, Yarnold, & August, in press; Vaughn et al., 1997). The former study established the relative value of the BASC PRS Attention Problem Scale for identifying students likely to demonstrate nonspecific ADHD, although both measures demonstrated limited ability to identify ADHD subtype. In the second study, both BASC parent and teacher scales demonstrated significant superiority over the CBCL and TRF in accurately classifying ADHD type.

The routine use of these measures will expand to repeated assessment paradigms. The interest

in measuring treatment effectiveness as a clinical as well as a research priority will make a meaningful contribution to identifying optimal treatment strategies as well as assigning specific treatment options for the individual case. These trends will facilitate retrospective "file drawer" study within clinical settings and expand the investigation of clinically referred youth.

4.13.4 REFERENCES

- Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs*, 80 (Whole Issue No. 615).
- Achenbach, T. M. (1981). A junior MMPI? (Review of multidimensional description of child personality: A manual for the Personality Inventory for Children and Actuarial assessment of child and adolescent personality: An interpretive guide for the Personality Inventory for Children profile.) *Journal of Personality Assessment*, 45, 332–333.
- Achenbach, T. M. (1991a). *Integrative guide for the 1991 CBCL/4-18, YSR, and TRF profiles*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1991b). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1991c). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1991d). *Manual for the Youth Self-Report and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. (1992). New developments in multi-axial empirically based assessment of child and adolescent psychopathology. In J. Rosen & P. McReynolds (Eds.), *Advances in psychological assessment* (Vol. 8, pp. 75–102). New York: Plenum.
- Achenbach, T. M. (1993). *Empirically based taxonomy: How to use syndromes and profile types derived from the CBCL/4-18, TRF, and YSR*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101, 213–232.
- August, G. J., MacDonald, A. W., Realmuto, G. M., & Skare, S. S. (1996). Hyperactive and aggressive pathways: Effects of demographic, family, and child characteristics on children's adaptive functioning. *Journal of Abnormal Child Psychology*, 25, 341–351.
- August, G. J., Realmuto, G. M., MacDonald III, A. W., Nugent, S. M., & Crosby, R. (1996). Prevalence of ADHD and comorbid disorders among elementary school children screened for disruptive behavior. *Journal of Abnormal Child Psychology*, 24, 571–595.
- Biederman, J., Faraone, S. V., Doyle, A., Lehman, B. K., Kraus, I., Perrin, J., & Tsuang, M. T. (1993). Convergence of the Child Behavior Checklist with structured interview-based psychiatric diagnoses of ADHD children with and without comorbidity. *Journal of Child Psychology and Psychiatry*, 34, 1241–1251.
- Bloomquist, M. L., August, G. J., Cohen, C., Doyle, A., & Everhart, K. (1997). Social problem solving in hyperactive-aggressive children: How and what they think in conditions of automatic and controlled processing. *Journal of Clinical Child Psychology*, 26, 172–180.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist scales for Attention-Deficit Hyperactivity Disorder: A receiver-operating characteristic analysis. *Journal of Consulting and Clinical Psychology*, 62, 1017–1025.
- Conners, C. K. (1997). *Conners' Rating Scales-Revised technical manual*. North Tonawanda, NY: Multi-Health Systems.
- Conners, C. K. (in press). Conners' Rating Scales-Revised. In M. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Conners, C. K., Wells, K. C., Parker, J. D. A., Sitarenios, G., Diamond, J. M., & Powell, J. W. (1997). A new self-report scale for assessment of adolescent psychopathology: Factor structure, reliability, validity, and diagnostic sensitivity. *Journal of Abnormal Child Psychology*, 25, 487–497.
- Dalton, J. E. (1996). Juvenile male sex offenders: Mean scores on the BASC Self-Report of Personality. *Psychological Reports*, 79, 634.
- Doyle, A., Ostrander, R., Skare, S., Crosby, & August, G. J. (1997). Convergent and criterion-related validity of the Behavior Assessment System for Children-Parent Rating Scale. *Journal of Clinical Child Psychology*, 26, 276–284.
- Drotar, D., Stein, R. E., & Perrin, E. C. (1995). Methodological issues in using the Child Behavior Checklist and its related instruments in clinical child psychology research. *Journal of Clinical Child Psychology*, 24, 184–192.
- Edelbrock, C., & Costello, A. J. (1988). Convergence between statistically derived behavior problem syndromes and child psychiatric diagnoses. *Journal of Abnormal Child Psychology*, 16, 219–231.
- Erford, B. T. (1996). Analysis of the Conners' Teacher Rating Scale-28 (CTRS-28). *Assessment*, 3, 27–36.
- Faraone, S. V., Biederman, J., Weber, W., & Russell, R. L. (1998). Psychiatric, neuropsychological, and psychosocial features of DSM-IV subtypes of Attention-Deficit/Hyperactivity Disorder: Results from a clinically referred sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37, 185–193.
- Flanagan, D. P., Alfonso, V. C., Primavera, L. H., & Povall, L. (1996). Convergent validity of the BASC and SSRS: Implications for social skills assessment. *Psychology in the Schools*, 33, 13–23.
- Forbes, G. B. (1985). The Personality Inventory for Children (PIC) and hyperactivity: Clinical utility and problems of generalizability. *Journal of Pediatric Psychology*, 10, 141–149.
- Gdowski, C. L., Lachar, D., & Kline, R. B. (1985). A PIC profile typology of children and adolescents: I. An empirically-derived alternative to traditional diagnosis. *Journal of Abnormal Psychology*, 94, 346–361.
- Greenbaum, P. E., Dedrick, R. F., Prange, M. E., & Friedman, R. M. (1994). Parent, teacher, and child ratings of problem behaviors of youngsters with serious emotional disturbances. *Psychological Assessment*, 6, 141–148.
- Jensen, P. S., Watanabe, H. K., Richters, J. E., Roper, M., Hibbs, E. D., Salzberg, A. D., & Liu, S. (1996). Scales, diagnoses, and child psychopathology: II. Comparing the CBCL and the DISC against external validators. *Journal of Abnormal Child Psychology*, 29, 151–168.
- Kamphaus, R. W., & Frick, P. J. (1996). *Clinical assessment of child and adolescent personality and behavior*. Boston, MA: Allyn & Bacon.

- Kamphaus, R. W., Huberty, C. J., DiStifano, C., & Petoskey, M. D. (1997). A typology of teacher-rated child behavior for a national U.S. sample. *Journal of Abnormal Child Psychology*, 25, 453-463.
- Kamphaus, R. W., & Reynolds, C. R. (1998). *BASC Monitor for ADHD manual*. Circle Pines, MN: American Guidance Service.
- Kline, R. B. (1994). New objective rating scales for child assessment, I. Parent- and teacher-informant inventories of the Behavior Assessment System for Children, the Child Behavior Checklist, and the Teacher Report Form. *Journal of Psychoeducational Assessment*, 12, 289-306.
- Kline, R. B. (1995). New objective rating scales for child assessment, II. Self-report scales for children and adolescents: Self-Report of Personality of the Behavior Assessment System for Children, the Youth Self-Report, and the Personality Inventory for Youth. *Journal of Psychoeducational Assessment*, 13, 169-193.
- Kline, R. B., & Lachar, D. (1992). Evaluation of age, sex, and race bias in the Personality Inventory for Children (PIC). *Psychological Assessment*, 4, 333-339.
- Kline, R. B., Lachar, D., & Gdowski, C. L. (1987). A PIC typology of children and adolescents: II. Classification rules and specific behavior correlates. *Journal of Clinical Child Psychology*, 16, 225-234.
- Kline, R. B., Lachar, D., Gruber, C. P., & Boersma, D. C. (1994). Identification of special education needs with the Personality Inventory for Children (PIC): A profile-matching strategy. *Assessment*, 1, 301-313.
- Kline, R. B., Lachar, D., & Sprague, D. J. (1985). The Personality Inventory for Children (PIC): An unbiased predictor of cognitive and academic status. *Journal of Pediatric Psychology*, 10, 461-477.
- Knoff, H. M. (1989). Review of the Personality Inventory for Children, Revised Format. In J. C. Connolly & J. C. Kramer (Eds.), *The tenth mental measurements yearbook* (pp. 624-630). Lincoln, NE: Buros Institute of Mental Measurements.
- Kovacs, M. (1992). *Children's Depression Inventory (CDI) manual*. Toronto: Multi-Health Systems.
- Lachar, D. (in press). Personality Inventory for Children-2nd Edition (PIC-2), Personality Inventory for Youth (PIY), and the Student Behavior Survey (SBS). In M. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Lachar, D. (1982). *Personality Inventory for Children (PIC) revised format manual supplement*. Los Angeles: Western Psychological Services.
- Lachar, D. (1993). Symptom checklists and personality inventories. In T. R. Kratochwill & R. J. Morris (Eds.), *Handbook of psychotherapy with children* (pp. 38-57). New York: Allyn & Bacon.
- Lachar, D. (1999). *Personality Inventory for Children-2nd Edition (PIC-2) manual*. Los Angeles: Western Psychological Services.
- Lachar, D., & Gdowski, C. L. (1979). *Actuarial assesment of child and adolescent personality: An interpretive guide for the Personality Inventory for Children profile*. Los Angeles: Western Psychological Services.
- Lachar, D., Gdowski, C. L., & Snyder, D. K. (1982). Broad-band dimensions of psychopathology: Factor scales for the Personality Inventory for Children. *Journal of Consulting and Clinical Psychology*, 50, 634-642.
- Lachar, D., & Gruber, C. P. (1993). Development of the Personality Inventory for Youth: A self-report companion to the Personality Inventory for Children. *Journal of Personality Assessment*, 61, 81-98.
- Lachar, D., & Gruber, C. P. (1995). *Personality Inventory for Youth (PIY) manual technical guide*. Los Angeles: Western Psychological Services.
- Lachar, D., Harper, R. A., Green, B. A., Morgan, S. T., & Wheeler, A. C. (1996, August). *The Personality Inventory for Youth: Contribution to diagnosis*. Paper presented at the 104th Annual Convention, American Psychological Association, Toronto, Canada.
- Lachar, D., & Kline, R. B. (1994). Personality Inventory for Children and Personality Inventory for Youth. In M. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment*. Hillsdale, NJ: Erlbaum.
- Lachar, D., & Kline, R. B., & Gdowski, C. L. (1987). Respondent psychopathology and interpretive accuracy of the Personality Inventory for Children: The evaluation of a "most reasonable" assumption. *Journal of Personality Assessment*, 51, 165-177.
- Lachar, D., Kline, R. B., Wingenfeld, S. A., & Gruber, C. P. (1999). *Student Behavior Survey (SBS) manual*. Los Angeles: Western Psychological Services.
- LaCombe, J. A., Kline, R. B., Lachar, D., Butkus, M., & Hillman, S. B. (1991). Case history correlates of a Personality Inventory for Children (PIC) profile typology. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 3, 678-687.
- Lett, N. J., & Kamphaus, R. W. (1997). Differential validity of the BASC Student Observation System and the BASC Teacher Rating Scale. *Canadian Journal of School Psychology*, 13, 1-14.
- Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology*, 19, 136-143.
- Lowman, M. G., Schwanz, K. A., & Kamphaus, R. W. (1996). WISC-III third factor: Critical measurement issues. *Canadian Journal of School Psychology*, 12, 15-22.
- Merrell, K. W. (1994). *Assessment of behavioral, social, & emotional problems. Direct & objective methods for use with children and adolescents*. New York: Longman.
- Michael, K. D., & Merrell, K. W. (1998). Reliability of children's self-reported internalizing symptoms over short to medium-length time intervals. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37, 194-201.
- O'Connor, T. G., McGuire, S., Reiss, D., Hetherington, E. M., & Plomin, R. (1998). Co-occurrence of depressive symptoms and antisocial behavior in adolescence: A common genetic liability. *Journal of Abnormal Psychology*, 107, 27-37.
- Ostrander, R., Weinfurt, K. P., Yarnold, P. R., & August, G. J. (in press). Diagnosing attention deficit disorders using the BASC and the CBCL: Test and construct validity analyses using optimal discriminant classification trees. *Journal of Consulting and Clinical Psychology*.
- Pearson, D. A., & Lachar, D. (1994). Using behavioral questionnaires to identify adaptive deficits in elementary school children. *Journal of School Psychology*, 32, 33-52.
- Phares, V. (1997). Accuracy of informants: Do parents think that mother knows best? *Journal of Abnormal Child Psychology*, 25, 165-171.
- Pisecco, S., Lachar, D., Gallen, R. T., Gruber, C. P., & Huzinec, C. (1998). *Development of disruptive behavior DSM-IV scales from teacher ratings*. Manuscript submitted for publication.
- Realmutto, G. M., August, G. J., Sieler, J. D., & Pessoa-Brandao (1997). Peer assessment of social reputation in community samples of disruptive and nondisruptive children: Utility of the revised class play method. *Journal of Clinical Child Psychology*, 26, 67-76.
- Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children manual*. Circle Pines, MN: American Guidance Service.
- Reynolds, C. R., & Richmond, B. O. (1985). *Revised Children's Manifest Anxiety Scale manual*. Los Angeles: Western Psychological Services.

- Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin*, 112, 485-499.
- Rowe, K. S., & Rowe, K. J. (1997). Norms for parental ratings on Conners' Abbreviated Parent-Teacher Questionnaire: Implications for the design of behavioral rating inventories and analyses of data derived from them. *Journal of Abnormal Child Psychology*, 25, 425-451.
- Steingard, R., Biederman, J., Doyle, A., & Sprich-Buckminster, S. (1992). Psychiatric comorbidity in attention deficit disorder: Impact on the interpretation of Child Behavior Checklist results. *Journal of the Academy of Child and Adolescent Psychiatry*, 31, 449-454.
- Vaughn, M., Black, K., Hall, J., Hynd, G., & Riccio, C. A. (1995). *Use of the BASC and Achenbach for assessment and intervention planning*. Paper presented at the 1995 Convention of the National Association of School Psychologists, Chicago.
- Vaughn, M. L., Riccio, C. A., Hynd, G. W., & Hall, J. (1997). Diagnosing ADHD (predominantly inattentive and combined subtypes): Discriminant validity of the Behavior Assessment System for Children and the Achenbach parent and teacher rating scales. *Journal of Clinical Child Psychology*, 26, 349-357.
- Vignoe, D., & Achenbach, T. M. (1997). *Bibliography of published studies using the Child Behavior Checklist and related materials: 1997 edition*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Wainwright, A., & MHS Staff (1996). *Conners' Rating Scales: Over 25 years of research—An annotated bibliography*. Toronto: Multi-Health Systems.
- Weinstein, S. R., Noam, G. G., Grimes, K., Stone, K., & Schwab-Stone, M. (1990). Convergence of DSM-III diagnoses and self-reported symptoms in child and adolescent inpatients. *Journal of the American Academy of Child and Adolescent Psychiatry*, 29, 627-634.
- Wingenfeld, S. A., Lachar, D., Gruber, C.P., & Kline, R. B. (1998). Development of the teacher-informant Student Behavior Survey. Manuscript submitted to *Journal of Psychoeducational Assessment*.

4.14

Objective Personality Assessment with Adults

JAMES N. BUTCHER and JEANETTE TAYLOR
University of Minnesota, Minneapolis, MN, USA

and

G. CYNTHIA FEKKEN
Queen's University, Kingston, ON, Canada

4.14.1 INTRODUCTION	404
4.14.1.1 <i>Intelligence and Insight</i>	404
4.14.1.2 <i>Veridicality of Self-report</i>	405
4.14.1.3 <i>Personality Stability</i>	405
4.14.1.3.1 <i>Definitions</i>	405
4.14.1.3.2 <i>Influences on personality stability</i>	406
4.14.2 ASSESSING ADULTS IN CLINICAL SETTINGS	407
4.14.2.1 <i>The Minnesota Multiphasic Personality Inventory-Revised</i>	407
4.14.2.1.1 <i>Origin of the MMPI/MMPI-2</i>	408
4.14.2.1.2 <i>Measurement dimensions</i>	408
4.14.2.1.3 <i>Recent validity research for the MMPI-2</i>	409
4.14.2.2 <i>Basic Personality Inventory</i>	413
4.14.2.3 <i>Personality Assessment Inventory</i>	415
4.14.3 SPECIALIZED OR FOCUSED CLINICAL ASSESSMENT MEASURES	416
4.14.3.1 <i>Millon Clinical Multiaxial Inventory</i>	416
4.14.3.2 <i>The Beck Depression Inventory</i>	418
4.14.3.3 <i>The State-Trait Anxiety Inventory</i>	418
4.14.3.4 <i>Whitaker Index of Schizophrenic Thinking</i>	419
4.14.4 NORMAL RANGE PERSONALITY ASSESSMENT	420
4.14.4.1 <i>Objective Personality Measures in Research</i>	420
4.14.4.1.1 <i>The Five Factor Model (the Big Five)</i>	420
4.14.4.1.2 <i>NEO Personality Inventory</i>	420
4.14.4.1.3 <i>Multidimensional Personality Questionnaire</i>	420
4.14.4.1.4 <i>Personality Research Form</i>	421
4.14.4.1.5 <i>Sixteen Personality Factor Test</i>	421
4.14.4.1.6 <i>California Psychological Inventory</i>	421
4.14.4.2 <i>Objective Personality Measures in Educational/Vocational Assessment</i>	422
4.14.4.2.1 <i>FFM</i>	422
4.14.4.2.2 <i>NEO-PI</i>	422
4.14.4.2.3 <i>CPI</i>	422
4.14.4.2.4 <i>MMPI/MMPI-2</i>	422
4.14.4.3 <i>Personnel Screening</i>	423
4.14.4.4 <i>Other Personality Measures in Personnel Selection</i>	423
4.14.4.5 <i>The 16PF</i>	423
4.14.4.6 <i>FFM</i>	423

4.14.4.7 NEO-PI	424
4.14.4.7.1 CPI	424
4.14.5 SUMMARY	424
4.14.6 REFERENCES	425

4.14.1 INTRODUCTION

Objective personality tests form a standard part of most applied psychologists' toolboxes when it comes to measuring personality. In contrast to projective techniques and to subjective approaches to personality assessment, objective personality questionnaires are made up of relatively unambiguous stimuli or items; offer the respondent relatively restricted response options; and present a scoring scheme that involves few, if any, scoring judgments, resulting in high scoring reliability (Wiggins, 1973). Most objective personality assessment instruments used by clinicians and researchers employ a self-report format. In contemporary psychology, the Minnesota Multiphasic Personality Inventory (MMPI) and the MMPI-2 are the most widely used personality assessment instruments in both clinical and research settings (Butcher and Rouse, 1996). The next most popular instruments used to assess personality, the Rorschach and the Thematic Apperception Test, are projective (not objective), but each still requires the clinician to obtain self-reported behavior. In essence, the clinician is dependent on the client's ability and willingness to make accurate self-reports when assessing personality with standardized instruments. It is, therefore, important to determine whether adults being assessed can competently reveal information about their personalities through self-report instruments. In order to address this question, the client's intelligence and insight in sharing self-information and the success of personality tests for eliciting self-information that can be externally verified need to be considered. We will examine each of these considerations below.

4.14.1.1 Intelligence and Insight

The existing literature on intelligence and personality is largely focused on the predictive relationship between the two constructs. However, clinicians are generally not concerned about the ability of a personality profile to predict an intelligence quotient or vice versa. Instead, most are concerned with the more basic question of how intellectual functioning impacts the ability of a client to self-disclose on an objective personality test. No literature exists which directly addresses the issue of the

relationship between a client's level of intelligence and his or her ability to competently reveal information about personality through a self-report instrument.

Objective personality measures leave the issue of the client's intellectual functioning in the hands of the clinician. None of the widely used personality assessment instruments require a pre-screening for normal or above normal intelligence. However, most self-report personality instruments require that the examinee have a minimum reading level competency (often expressed as a minimum grade level). Thus, a minimum level of school achievement is implicit in the validity of a client's score, and it follows that the clinician should use care when administering a self-report instrument to a client with known or suspected intellectual impairment. It is also essential for the clinician to be aware of the educational range of the standardization sample for any self-report measure, and to use caution when interpreting protocols from clients who fall outside of that range.

The clinician needs to be aware of a client's intellectual functioning as it may influence the client's ability to provide information about his or her personality on a self-report measure. If a client appears to be low in intellectual functioning, then the results of self-report measures should be interpreted with caution and perhaps supplemented with reports from other sources.

While the capacity for insight cannot readily be extrapolated from an intelligence quotient in a linear fashion, insight into one's problems might be discernible from the personality profile itself. As Butcher and Rouse (1996) note in their recent review of personality research, several of the widely used objective personality measures have attitudinal measures that aid the clinician in evaluating the client's disclosure capability. Crookes and Buckley (1976) found a relationship between the Eysenck Personality Inventory Lie (L) scale score and diagnosis among both psychiatric inpatients and outpatients. They found that high L scale scores were positively related to diagnosis of disorders associated with low insight. They concluded that the L score (a validity indicator found on several objective personality instruments) is highly related to a person's awareness of his or her behavior. Moreover, this finding is in agreement with the interpretive guidelines for the MMPI-2 L scale, which suggests the general

utility of this measure at appraising a client's self-awareness.

The quality of insight that a client brings to the assessment situation may greatly influence the validity of the assessment. Some instruments are designed to assess clients through behavioral items that require low insight or self-observation. These items are specifically selected because of their ability to predict relevant criteria and may appear at face value totally unrelated to the particular aspects of personality functioning intended to be predicted. This highlights one advantage of using objective personality measures in clinical settings: clients who may yield poor information during a clinical interview that requires self-evaluation and/or introspection may be fully capable of providing valuable clinical information when a personality instrument comprising empirically selected items is employed.

4.14.1.2 Veridicality of Self-report

Aside from the issues of intellectual functioning and insight the utility of self-report measures can be assessed by examining their success at corroborating independent facts about a client. For example, in an early study of the MMPI, Payne and Wiggins (1972) examined the relationship between content-based profiles and external descriptors of a large group of psychiatric inpatients. Interpretation of content-based profiles is based on combinations of "obvious" test items which correspond to traditional self-report instruments (e.g., "I am a high-strung person"). That is, the client endorses the item as a means of directly relaying information about him- or herself. The authors found that the patients' self-reported MMPI profiles matched quite well with interview report and external observation.

In a similar vein, Koss and Butcher (1973) found that psychiatric inpatients identified (through their observed behavior and presenting symptoms) as belonging to one of six operationally defined crisis situations could be discriminated from one another on the basis of their endorsed MMPI content. The authors interpreted their results as providing evidence for the competence and willingness of adults in clinical settings to reflect information accurately about their personality and psychological functioning through self-report test items. Similar evidence of a positive relationship between self-reported personality measures and external criteria has been found for certain scales from the California Psychological Inventory (CPI; e.g., Hindelang, 1972), the Millon Clinical Multiaxial Inventory (MCMI; e.g.,

Jaffe and Archer, 1987), and the Basic Personality Inventory (BPI; Holden, Fekken, Reddon, Helmes, & Jackson, 1988), to name three.

Although the research literature shows that people can competently disclose information about their personalities through standardized, objective, self-report measures, some individuals, however, may not be motivated to cooperate in their psychological assessment. That is, some respondents, because of a need or motivation to present themselves in a particular way, do not respond in a truthful, open manner. The most common examples of individuals being motivated to appear different from the way they actually are in today's assessment settings are applicants for employment, parents involved in custody disputes, and other individuals being evaluated as part of a court case. A self-report assessment instrument must contain an effective means of identifying respondents who are dissimulating. Perhaps the most common way of detecting whether people are not motivated to report accurately about themselves is via validity scales or control scales. In some assessment situations the validity scales provide the most important and useful information about the client. A self-report personality measure without effective validity scales is too limited to operate across a broad range of assessment situations. Furthermore, tests that possess validity indicators, such as the L scale on the MMPI/MMPI-2, offer the clinician information regarding both the client's willingness to cooperate with the assessment and his or her level of insight. These points argue in favor of the continued use of standardized, objective self-report measures in the assessment of personality in clinical settings.

4.14.1.3 Personality Stability

One of the assumptions of objective personality assessment is that the personality characteristic being measured is stable. Various major studies have come to the conclusion that personality changes little in adulthood, particularly after age 30 (Conley, 1985; Finn, 1986; McCrae & Costa, 1990; Schuerger, Zarrella, & Hotz, 1989). Typical levels of stability on personality tests are in the 0.5–0.7 range. This generalization bears some comment.

4.14.1.3.1 Definitions

What exactly is meant by stable responding? On an objective personality questionnaire, a set of responses that are consistent over time may be defined in at least three ways: (i) as an identical scale score, (ii) as a set of identical item

responses, or (iii) as a scale score that signifies the same relative standing on the personality characteristic.

Conclusions about the stability of personality are most commonly based on scale scores. That is, the stability of personality estimates result from comparing scale scores obtained at two points in time. Presumably the same scale score could result even if the test respondent endorsed somewhat different subsets of items. The consistency of responses to individual test items has been studied as a meaningful individual differences variable and as an index for establishing the interpretability of individual response protocols. Item response consistency has not been examined as a definition of longitudinal personality stability *per se*.

Many researchers have focused on changes in average scale scores. Their goal is to understand normative changes in personality. On the other hand, some researchers examine personality stability by correlating scale scores. Thus, when coefficients of personality stability are reported to be in the 0.5–0.7 range, this should be understood to be a statement about relative, and not absolute, scale score stability.

4.14.1.3.2 Influences on personality stability

Although the typical stability of personality scale scores is quite high, there are nonetheless particular variables that affect score stability in predictable ways. The first consideration is the influence of instrument characteristics. Specifically, the average number of items per scale and the homogeneity of the scales are the important predictors of scale score stability (Finn, 1986; Schuerger et al., 1989). Variance associated with specific instruments (e.g., the MMPI vs. the Sixteen Personality Factor Test [16PF]) apparently does not contribute to greater personality stability beyond the effects of scale length and homogeneity (Schuerger et al., 1989). Such findings underscore the need for test users to select reliable measures, particularly if test data are intended to evaluate long-term personality change.

Second, estimates of personality stability are affected by the length of the retest interval. Not surprisingly, personality appears to more stable over short intervals than long ones. In particular, score stability drops over the first year before it levels off (Schuerger et al., 1989; Windle, 1954) and then stays stable at the same level for very long periods of time (McCrae & Costa, 1990).

One simple reason for relatively high short-term personality stability may be that people remember and repeat their responses to the items on a questionnaire. Lumsden (1977),

however, proposed a substantive rather than a methodological explanation. He argued that particular personality characteristics “swell” into prominence over the short term and then fade again. For a student entering college, for example, “independence” may become a salient dimension and hence, self-reports on this dimension are likely to have exaggerated consistency over the (short) time period when “independence” is prominent.

A third influence on personality stability has to do with the operationalization of personality. Most objective personality questionnaires are intended to measure traits that are by definition stable and enduring. The construction of such questionnaires favors selection of items that will yield stable scores. In addition, the overall instructions as well as the wording of specific items prompt people to make broad generalizations about themselves and to downplay variations. Thus, personality may appear to be stable because the constructs of personality are explicitly conceptualized and measured as stable. There are certainly empirical studies that challenge the idea of personality stability by providing evidence of developmental change over long time periods based on personality constructs and measures that expressly incorporate the notion of change (e.g., Whitbourne, Zuschlag, Elliot, & Waterman, 1992).

Fourth, some personality constructs are associated with more stability than others. Schuerger et al.’s (1989) data show that measures of psychopathology have less scale score stability than measures of “normal” personality characteristics. Perhaps “normal” personality constructs are more crystallized than psychopathological constructs and thus less error is built into the construct itself. Similarly, psychopathological constructs may be more confounded with the effects of response style variance such as social desirability and acquiescence than normal personality constructs, again compounding the noisiness of the construct.

Many researchers have tried to highlight differences in stability among the specific constructs that they chose to study. For example, Finn (1986) demonstrated that mood-related constructs (e.g., depression) had comparatively low levels of consistency, whereas Helson and Moane (1987) found that constructs related to socialization had high stability. There are shortcomings to such an approach. The identification of those dimensions on which people are changeable (i.e., states) as distinct from the more permanent individual differences dimensions needs to be tackled systematically if a comprehensive list is to be obtained (Cattell, 1963).

The interpretation of differences in stability between constructs is further complicated by age and cohort effects. Studies such as that of Helson and Moane (1987) estimate personality stability using information obtained from a single group of subjects who have been assessed at different points in time. There are limits to the generalizability of this information to other groups of people. Other studies such as that of Finn (1986) use basically a cross-sectional approach. They assess two or more groups of subjects having different ages at a single time. Here it is difficult to attribute differences in the stability of particular personality constructs to age as opposed to generational factors. More complicated designs may be needed in order to understand such age and cohort effects (Conley, 1985). At the very least, any generalizations about which personality constructs are more or less stable need to be carefully qualified.

A fifth influence on personality stability has to do with features of the persons being evaluated. Test respondents exhibit remarkable individual differences in personality stability (Assendorp, 1992; McCrae & Costa, 1990). Do some subgroups of people show more personality stability? Across studies, patient and prisoner samples show less personality test score stability than "normals" (Schuerger et al., 1989). Perhaps "non-normals" have systematically higher scale scores than "normals" and, hence, their retest scores show regression toward the mean (Windle, 1954). Alternatively, "non-normal" groups also may be more likely than normal groups to seek actively to change personality.

The other two person variables that have received attention in the literature on personality stability are age and sex differences. Some researchers report that older adults are more stable than younger adults on many traits (Finn, 1986; Schuerger et al., 1989) although others argue that stability differences after about age 30 are minimal (McCrae & Costa, 1990). Finn (1986) has asked whether older adults might just be more rigid in their self-perceptions. However, the picture of personality stability that results from self-reports does tend to be substantiated by ratings on personality information collected from others, such as spouses (McCrae & Costa, 1990). Thus, personality may indeed be more stable in older adults.

Many studies in the area of personality stability sample exclusively men or women. Yet there seems to be little evidence of an overall difference in the level of personality stability for men and women (Schuerger et al., 1989). Nonetheless, researchers interested in normative change often argue that men (Finn, 1986) and women (Helson & Moane, 1987) may be

expected to change on different constructs as a function of distinct life paths or societal pressures. Few studies have been explicitly designed to compare sex differences in personality stability.

Overall, personality shows considerable stability in adulthood. The objective measures traditionally used in personality assessment may predispose us to find evidence of stability as a function of the kinds of constructs, instructions, and test construction techniques associated with these measures. Aspects of the test instrument itself, such as length or homogeneity, may moderate stability estimates, as could the length of the interval between personality assessments. Samples of "normals" and of older adults may exhibit more personality stability; sex differences in general personality stability appear to be minimal. The assumption of personality stability, at least over the short run, is central to objective personality assessment if test scores are going to predict relevant behaviors and have an impact on planning how to manage that behavior.

4.14.2 ASSESSING ADULTS IN CLINICAL SETTINGS

In this section we will provide an overview of a number of objective personality inventories that have been designed to assess adults in clinical settings. More detail on the use of the MMPI-2 will be given because it is the most widely researched and used measure. Two other measures that have recently been published to measure essentially the same clinical problem areas as the MMPI-2, BPI (Jackson, 1989) and the Personality Assessment Inventory (PAI) (Morey, 1991) will be briefly described. In the next section, several specialized measures will then be surveyed that have been developed to measure more specific clinical problems or behaviors: the MCMI to assess personality disorders; the Beck Depression Inventory (BDI; Beck, Steer, & Garbin, 1988) to assess anxiety; the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970); and the Whitaker Index of Schizophrenic Thinking (WIST; Whitaker, 1973).

4.14.2.1 The Minnesota Multiphasic Personality Inventory-Revised

The MMPI-2 is the most widely used personality test with adults (Lubin, Larsen, & Matarazzo, 1984; Watkins, 1996). Keilen and Bloom (1986) found that the MMPI/MMPI-2 was the most frequently used test in custody evaluations. Developed originally in the late

1930s by Starke Hathaway and J. C. McKinley and redeveloped by Butcher, Dahlstrom, Graham, Tellegen, and Kaemmer (1989), this instrument provides a comprehensive survey of personality characteristics and clinical problems. In the original MMPI, a strictly empirical scale construction approach was followed to develop scales that would assess a patient's probable "membership" in a clinical group. An extensive amount of research has been published on the effectiveness of these measures at predicting and describing problems in adults. Research on the original version of the MMPI covered a very broad range of people—psychiatric and medical patients, substance abusers, incarcerated felons, and many other clinical groups.

In addition, a very broad range of "normals" have been studied including applicants for various jobs such as airline pilots, US Navy submariner crew members, and police and security personnel. Moreover, the instrument came to be widely employed as a personality research instrument. Butcher and Rouse (1996), in a survey of 20 years of research in clinical assessment, found that the MMPI/MMPI-2 is the most widely researched instrument with nearly twice the number of articles as the seven next leading tests. What has made the MMPI/MMPI-2 the most widely used assessment technique in the personality area? We will examine the make-up, utility, and limitations of the instrument in assessing adults.

4.14.2.1.1 Origin of the MMPI/MMPI-2

A growing dissatisfaction with subjective methods of evaluating patients in clinical situations (such as interviews and projective tests) led the original MMPI developers, Hathaway (a psychologist) and McKinley (a psychiatrist), to experiment with an objective method of clinical problem assessment. They accumulated a large number of items (symptoms, beliefs, attitudes, etc.) from tests and case material and administered them to a large group of "normal" people to serve as a comparison group. They then administered the items to well-defined and homogeneous groups of clinical patients. They empirically contrasted the responses of the patient groups with the normals to obtain items that significantly separated the groups. These items were then combined into scales for the various clinical problem areas such as depression, schizophrenia, and so forth. Once derived, the clinical scales were normed on the non-patient sample and *T* scores were developed to enable the test interpreter to determine how extreme a particular person's score was on a given scale. All of the scales were normed so that

they had a mean score of 50 and a standard deviation of 10. These standard score distributions then allowed the scores to be plotted on a profile so that the interpreter would have a visual picture of how extreme a particular score was when compared with the normal. Their empirical approach produced highly valid and effective scales that predicted or described the likelihood that a person's score on a scale was in the clinical range or similar to the patient groups.

The original MMPI clinical scales have undergone very substantial study and cross-validation since their publication. They have become a standard means of objective symptom classification since their development. The clinical scales and configurations of scales, referred to as code types, have undergone substantial documentation as an objective classification schema—often referred to as "cookbooks." Researchers, for example, have cataloged the behavioral characteristics associated with their test indices allowing for automatic interpretation. That is, when a particular scale score or cluster of scales are obtained, then a well-validated set of behavioral descriptions—known as descriptors—are applied. Their objective classification approach to interpretation fostered the development of computer-based interpretation methods so popular today (Butcher, 1995; Butcher et al., 1998).

4.14.2.1.2 Measurement dimensions

(i) Validity scales

As noted earlier, it is essential to any self-reported personality assessment to appraise carefully possible invalidating conditions or circumstances. Structural elements of the test administration require evaluation to determine if extratest factors influenced the item responses. For example, were the instructions clearly presented, was the individual able to read and comprehend the items?

The MMPI-2 contains a number of scales and indexes that provide the test interpreter with information on the individual's cooperativeness and honesty in responding to the items (see Table 1 for a summary of the validity or control scales for the MMPI-2). A clear picture is obtained as to whether the person has attempted to present a false picture on the test. These scales are of several types. First, noncontent-oriented measures can provide information on whether the person was inconsistent in responding (the True Response Inconsistency or TRIN and Variable Response Inconsistency or VRIN scales). Additionally, the Cannot Say or "?" scale provides information about the person's cooperation in completing the items. Two other

indexes that can provide clues to uncooperative test-taking behavior are the percentage of true and percentage of false endorsement. Records with a nearly all true or all false response pattern suggest uncooperativeness in completing the task.

Three scales have been developed to provide information about the tendency on the part of some people to claim excessive virtue or to deny problems. The L scale assesses an unsophisticated tendency to claim excessive virtue; the K scale measures test defensiveness; and the S scale (Superlative Self-presentation) assesses the tendency of some people to present themselves in a highly favorable manner (Butcher & Han, 1995).

Three scales have been developed to assess the tendency that some respondents have to claim excessive mental health symptoms. The original F scale developed by Hathaway and McKinley (1940) assesses excessive symptom checking by attending to extreme or rare responses. Those who endorse a large number of rare items are thought to be presenting an unselective or exaggerated complaint pattern. High scores on F have been associated with malingering (Berry, Baer, & Harris, 1991; Schretlen, 1988).

The F(b) scale, an infrequency scale that covers extreme items in the back of the MMPI-2 items operates much like the original F scale (Butcher et al., 1989). The newest validity measure developed by Arbisi and Ben-Porath (1995) is an extremely valuable type of infrequency scale that differs from the original F scale in an important way: the F(p) scale assesses extreme responding in a clinical sample. The F(p) scale assesses infrequent responding in a clinical setting and thereby, when extreme, suggests malingering of psychiatric symptoms.

(ii) Clinical scales

The traditional MMPI clinical scales (Hathaway & McKinley, 1940) have been updated in the MMPI-2 (Butcher et al., 1989) and have been extensively revalidated in their revised form (Archer, Griffin, & Aiduk, 1995; Butcher & Williams, 1992; Graham & Ben-Porath, 1990; Graham & Butcher, 1988). The MMPI-2 clinical scales are empirically derived measures of several well-established clinical patterns such as hypochondriasis, depression, paranoia, and schizophrenia (see Table 2).

These scales have been well studied over more than 50 years. They were kept nearly intact in MMPI-2 because of their descriptive power and ability to generalize across groups.

Two scales on the clinical profile, *Mf* and *Si*, are not empirically derived scales as were the original clinical scales but were included on the

instrument to assess personality characteristics of masculine–feminine interests and social introversion.

(iii) Content scales

The expanded item pool of the MMPI-2 allowed for the development of a new set of content-based scales to assess an expanded array of problems compared with the traditional clinical measures. The 15 content scales (see Table 3) were developed according to a multi-method/multistage strategy. The 567 items were subjected to a content analysis in order to derive homogeneous content groups rationally. This approach is not as theoretically blind as it seems, since the original MMPI content had been well delineated by Wiggins (1969) and many of these constructs were still available in a modified or reduced form in the revised instruments. Moreover, the MMPI-2 committee also wrote new items to cover contents that were limited or not available in MMPI-2, for example, suicide ideation, substance abuse, and type A behavior. It therefore became possible to develop scales for a number of problem areas that have clinical relevance and are sufficiently large to assess reliably. Following provisional derivation, other scale construction strategies such as internal consistency, item scale correlation, and external validation were followed to define and improve the item groups.

Content scales provide a different type of information to the clinical interpreter's resources. These scales are viewed as direct communications between the client and the clinician. The obvious content on the scales enables the client to address problems he or she has that are considered important to address in the clinical intervention.

In addition to their value as summaries or themes considered pertinent by the patient the content scales have clearly established external validity. A number of studies have provided data on the external validity of the content scales (Ben-Porath, Butcher, & Graham, 1991; Butcher, Graham, Williams, & Ben-Porath, 1990).

4.14.2.1.3 Recent validity research for the MMPI-2

An empirically derived and based instrument requires substantial research validation in order to be useful. The MMPI-2 has been substantially researched and validated. The fact that the MMPI-2 clinical scales are the same items as in the original instrument means that all of the research on the traditional measures applies with the revised form.

Table 1 Personality characteristics associated with validity indicator elevations.

<i>? Cannot say score</i> The total number of unanswered items. A defensive protocol with possible attenuation of scale scores is suggested if the ? raw score is more than 35.
<i>L (Lie) scale</i> A measure of an unsophisticated or self-consciously “virtuous” test-taking attitude. Elevated scores (above 70 <i>T</i>) suggest that the individual is presenting himself or herself in an overly positive light, attempting to create an unrealistically favorable view of his or her adjustment.
<i>F (Infrequency) scale</i> A high score (<i>T</i> above 90) suggests an exaggerated pattern of symptom checking that is inconsistent with accurate self-appraisal and suggests confusion, disorganization, or actual faking of mental illness. <i>T</i> scores above 100 invalidate the profile.
<i>F(B) scale</i> A second infrequency scale that appears toward the back of the MMPI-2 item pool is used to assess exaggerated responding at the end of the test. This scale operates much like the original MMPI F scale in detecting malingering, random responding, or motivation to exaggerate symptoms.
<i>F(p) scale</i> The Psychopathology Infrequency Scale F(p) was developed by Arbisi and Ben-Porath (1995) to assess infrequent responding in psychiatric settings. This scale is valuable in appraising the tendency for some people to exaggerate mental health symptoms in the context of patients with genuine psychological disorder.
<i>K (Defensiveness) scale</i> Measures an individual’s willingness to disclose personal information and discuss his or her problems. High <i>K</i> scores (<i>T</i> above 65) reflect an uncooperative attitude and an unwillingness or reluctance to disclose personal information. Low scores (<i>T</i> below 45) suggest openness and frankness. This scale is positively correlated with intelligence and educational level, which should be taken into account when interpreting the scores.
<i>Variable response inconsistency</i> The VRIN scale consists of 49 pairs of specially selected items. The members of each VRIN item pair have either similar or opposite content; each pair is scored for the occurrence of an inconsistency in responses to the two items. The scale score is the total number of item pairs answered inconsistently. High VRIN scores are a warning that a test subject may have been answering the items in the inventory in an indiscriminate manner, and raise the possibility that the protocol may be invalid and that the profile is essentially uninterpretable.
<i>True response inconsistency</i> The TRIN is made up of 20 pairs of items that are opposite in content. If a subject responds inconsistently by answering True to both items of certain pairs, one point is added to the TRIN score; if the subject responds inconsistently by answering False to certain item pairs, one point is subtracted. A very high TRIN score indicates a tendency to give True answers to the items indiscriminately (“acquiescence”), and a very low TRIN score indicates a tendency to answer False indiscriminately (“nonacquiescence”). (Negative TRIN scores are avoided by adding a constant to the raw score.) Very low or very high TRIN scores are a warning that the test subject may have been answering the inventory indiscriminately so that the profile may be invalid and uninterpretable.
<i>S scale</i> A new MMPI-2 scale was developed in an effort to improve the assessment of highly virtuous responding. The subjects used in the development of the scale were 274 male airline applicants and the MMPI-2 normative sample (<i>N</i> = 1138 men and 1462 women). The S scale was initially developed by examining item response differences between airline pilot applicants, who tend to engage in superlative self-description in order to impress examiners, and normative men from the MMPI-2 restandardization sample. The scale was refined by using internal consistency methods to ensure high-scale homogeneity. A factor analysis of the resulting 50 item scale (<i>S</i>) yielded five factors named: Beliefs in “Human Goodness”; Serenity; Contentment with Life; Patience and Denial of Irritability and Anger; and Denial of Moral Flaws. Linear <i>T</i> score conversion tables were computed for both men and women separately and combined (unisex) using the MMPI-2 restandardization data sets. The S scale was shown to have a number of behavioral correlates reflecting the presentation of oneself as a well-controlled, problem-free person.

Table 2 Empirical correlates for MMPI-2 clinical scales.*Scale 1 (Hypochondriasis)*

High-scoring people show: excessive bodily concern; somatic symptoms that tend to be vague and undefined; fatigue, pain, weakness; manifest anxiety; selfish, self-centered, and narcissistic behavior; pessimistic, defeatist, cynical outlook on life; dissatisfied and unhappy; make others miserable with whining, complaining behavior; demanding and critical of others; expresses hostility indirectly; rarely act out; dull, unenthusiastic, unambitious; ineffective in oral expression, longstanding health concerns; function at a reduced level of efficiency without major incapacity; not very responsive to therapy, tend to terminate therapy when therapist is seen as not giving enough attention and support; tend to seek medical solutions to life problems.

Scale 2 (Depression)

High-scoring people show: depressed, unhappy, and dysphoric behavior; they are pessimistic and self-deprecating; tend to feel guilty; report being sluggish; have somatic complaints such as weakness, fatigue, and loss of energy; are agitated, tense, highly-strung, and irritable; prone to worry; lack self-confidence; feel useless and unable to function; feels like a failure at school or on the job; introverted, shy, retiring, timid, and seclusive; aloof, maintain psychological distance; avoids interpersonal involvement; cautious, and conventional; has difficulty making decisions; nonaggressive; overcontrolled, deny impulses; make concessions to avoid conflict; motivated for therapy.

Scale 3 (Hysteria)

High-scoring people show: poor response to stress; they avoid responsibility through development of physical symptoms; have headaches, chest pains, weakness, and tachycardia, anxiety attacks; symptoms appear and disappear suddenly; lack insight about causes of symptoms; lack insight about own motives and feelings; lack anxiety, tension, and depression; rarely report delusions, hallucinations, or suspiciousness; psychologically immature, childish, and infantile; self-centered, narcissistic, and egocentric; expect attention and affection from others; use indirect and devious means to get attention and affection; do not express hostility and resentment openly; socially involved; friendly, talkative, and enthusiastic; superficial and immature in interpersonal relationships; show interest in others for selfish reasons; occasionally act out in sexual or aggressive manner with little apparent insight; initially enthusiastic about treatment; respond well to direct advice or suggestion; slow to gain insight into causes of own behavior; resistant to psychological interpretations.

Scale 4 (Psychopathic deviate)

High-scoring people show: antisocial behavior; rebellious toward authority figures; stormy family relationships; blame parents for problems; history of under-achievement in school; poor work history; marital problems; impulsive; strive for immediate gratification of impulses; do not plan well; act without considering consequences of actions; impatient; limited frustration tolerance; poor judgment; take risks; do not profit from experience; immature, childish, narcissistic, self-centered, and selfish; ostentatious, exhibitionistic; insensitive; interested in others in terms of *how* they can be used; likeable and usually create a good first impression; shallow, superficial relationships, unable to form warm attachments; extroverted, outgoing; talkative, active, energetic, and spontaneous; intelligent; assert self-confidence; have wide range of interests; lack definite goals; hostile, aggressive; sarcastic, cynical; resentful, rebellious; act out; antagonistic; aggressive outbursts, assaultive behavior; little guilt over negative behavior; may feign guilt and remorse when in trouble; free from disabling anxiety, depression, and psychotic symptoms; likely to have personality disorder diagnosis (antisocial or passive-aggressive); prone to worry; dissatisfied; show absence of deep emotional response; feel bored and empty; poor prognosis for change in therapy; blame others for problems; intellectualize; may agree to treatment to avoid jail or some other unpleasant experience but are likely to terminate before change is effected.

*Scale 5 (Masculinity-femininity)***MALES**

Scores of T over 80: Show conflicts about sexual identity; insecure in masculine role; effeminate; aesthetic and artistic interests; intelligent and capable; value cognitive pursuits; ambitious, competitive, and persevering; clever, clear-thinking, organized, logical; show good judgment and common sense; curious; creative, imaginative, and individualistic in approach to problems; sociable; sensitive to others; tolerant; capable of expressing warm feelings toward others; passive, dependent, and submissive; peace-loving; make concessions to avoid confrontations; good self-control; rarely act out. (The interpretation of high 5 scores should be tempered for males with advanced academic degrees.)

High T score between 70 and 79: May be viewed as sensitive; insightful; tolerant; effeminate; showing broad cultural interests; submissive, passive. (In clinical settings, the patient might show sex role confusion; or heterosexual adjustment problems.)

Table 2 (continued)

Low $T < 35$: "Macho" self-image, present self as extremely masculine; overemphasize strength and physical prowess; aggressive, thrill-seeking, adventurous, and reckless; coarse, crude, and vulgar; harbor doubts about own masculinity; have limited intellectual ability; narrow range of interests; inflexible and unoriginal approach to problems; prefer action to thought; are practical and nontheoretical; easy-going, leisurely and relaxed; cheerful, jolly, humorous; contented; willing to settle down; unaware of social stimulus value; lack insight into own motives; unsophisticated.

FEMALES

Scores of T over 70: Reject traditional female roles and activities; masculine interests in work, sports, hobbies; active, vigorous, and assertive; competitive, aggressive, and dominating; coarse, rough, and tough; outgoing, uninhibited, and self-confident; easy-going, relaxed, balanced; logical, calculated; unemotional, and unfriendly.

Low $T < 35$: Describe self in terms of stereotyped female role; doubts about own femininity; passive, submissive, and yielding; defer to males in decision-making; self-pity; complaining, fault finding; constricted; sensitive; modest; idealistic. (This interpretation for low 5 females does *not* apply for females with postgraduate degrees.)

Scale 6 (Paranoia)

Extremely high-scoring people ($T > 80$) show: frankly psychotic behavior; disturbed thinking; delusions of persecution and/or grandeur; ideas of reference; feel mistreated and picked on; angry and resentful; harbor grudges; use projection as defense; most frequently diagnosed as schizophrenia or paranoid state.

Moderate elevations (= 65–79 for males; 71–79 for females): Paranoid predisposition; sensitive; overly responsive to reactions of others; feel they are getting a raw deal from life; rationalize and blame others; suspicious and guarded; hostile, resentful, and argumentative; moralistic and rigid; overemphasizes rationality; poor prognosis for therapy; do not like to talk about emotional problems; difficulty in establishing rapport with therapist.

Extremely low ($T < 35$): should be interpreted with caution. In a clinical setting, low 6 scores, in the context of a defensive response set, may suggest frankly psychotic disorder; delusions, suspiciousness, ideas of reference; symptoms less obvious than for high scorers; evasive, defensive, guarded; shy, secretive, withdrawn.

Scale 7 (Psychasthenia)

High-scoring people show: anxious, tense, and agitated; high discomfort; worried and apprehensive; high strung and jumpy; difficulties in concentrating; introspective, ruminative; obsessive, and compulsive; feel insecure and inferior; lack self-confidence; self-doubting, self-critical, self-conscious, and self-derogatory; rigid and moralistic; maintain high standards for self and others; overly perfectionistic and conscientious; guilty and depressed; neat, orderly, organized, and meticulous; persistent; reliable; lack ingenuity and originality in problem solving; dull and formal; vacillates; are indecisive; distort importance of problems, overreact; shy; do not interact well socially; hard to get to know; worry about popularity and acceptance; sensitive, physical complaints; shows some insight into problems; intellectualize and rationalize resistant to interpretations in therapy; express hostility toward therapist; remain in therapy longer than most patients; makes slow but steady progress in therapy.

Scale 8 (Schizophrenia)

Very high scorers (= over 80–90) show: blatantly psychotic behavior; confused, disorganized, and disoriented; unusual thoughts or attitudes; delusions; hallucinations; poor judgment.

High (65–79): schizoid lifestyle; do not feel a part of social environment; feel isolated, alienated, and misunderstood; feel unaccepted by peers; withdrawn, seclusive, secretive, and inaccessible; avoid dealing with people and new situations; shy, aloof, and uninvolved; experience generalized anxiety; resentful, hostile, and aggressive; unable to express feelings; react to stress by withdrawing into fantasy and daydreaming; difficulty separating reality and fantasy; self-doubts; feel inferior, incompetent, and dissatisfied; sexual preoccupation, and sex role confusion; nonconforming, unusual, unconventional, and eccentric; vague, long-standing physical complaints; stubborn, moody, and opinionated; immature, and impulsive; highly-strung; imaginative; abstract, vague goals; lack basic information for problem-solving; poor prognosis for therapy; reluctant to relate in meaningful way to therapist; stay in therapy longer than most patients; may eventually come to trust therapist.

Scale 9 (Hypomania)

High-scoring people ($T > 80$) show: overactivity; accelerated speech; may have hallucinations or delusions of grandeur; energetic and talkative; prefer action to thought; wide range of interest; do not utilize energy wisely; do not see projects through to completion; creative, enterprising, and ingenious; little interest in routine or detail; easily bored and restless; low frustration tolerance; difficulty in inhibiting expression of impulses; episodes of irritability, hostility, and aggressive outbursts; unrealistic, unqualified optimism; grandiose aspirations; exaggerates self-worth and self-importance; unable to see own limitations; outgoing, sociable, and

Table 2 (continued)

gregarious; like to be around other people; create good first impression; friendly, pleasant, and enthusiastic; poised, self-confident; superficial relationships; manipulative, deceptive, unreliable; feelings of dissatisfaction; agitated; may have periodic episodes of depression; difficulties at school or work, resistant to interpretations in therapy; attend therapy irregularly; may terminate therapy prematurely; repeat problems in stereotyped manner; not likely to become dependent on therapists; becomes hostile and aggressive toward therapist.

Moderately elevated scores ($T > 65$, $LE\ 79$): Over-activity; exaggerated sense of self-worth; energetic and talkative; prefer action to thought; wide range of interest; do not utilize energy wisely; do not see projects through to completion; enterprising, and ingenious; lack interest in routine matters; become bored and restless easily; low frustration tolerance; impulsive; has episodes of irritability, hostility, and aggressive outbursts; unrealistic, overly optimistic at times; shows some grandiose aspirations; unable to see own limitations; outgoing, sociable, and gregarious; like to be around other people; create good first impression; friendly, pleasant, and enthusiastic; poised, self-confident; superficial relationships; manipulative, deceptive, unreliable; feelings of dissatisfaction; agitated; view therapy as unnecessary; resistant to interpretations in therapy; attend therapy irregularly; may terminate therapy prematurely; repeat problems in stereotyped manner; not likely to become dependent on therapists; become hostile and aggressive toward therapist.

Low scorers (T below 35): Low energy level; low activity level; lethargic, listless, apathetic, and phlegmatic; difficult to motivate; report chronic fatigue, physical exhaustion; depressed, anxious, and tense; reliable, responsible, and dependable; approach problems in conventional, practical, and reasonable way; lack self-confidence; sincere, quiet, modest, withdrawn, seclusive; unpopular; overcontrolled; unlikely to express feelings openly.

Scale 10 (Social introversion)

High-scoring people (> 65) show: socially introversion more comfortable alone or with a few close friends; reserved, shy, and retiring; uncomfortable around members of opposite sex; hard to get to know; sensitive to what others think; troubled by lack of involvement with other people; overcontrolled; not likely to display feelings openly; submissive and compliant; overly accepting of authority; serious, slow personal tempo; reliable, dependable; cautious, conventional, unoriginal in approach to problems; rigid, inflexible in attitudes and opinions; difficulty making even minor decisions; enjoys work; gain pleasure from productive personal achievement; tend to worry; are irritable and anxious; moody, experience guilt feelings; have episodes of depression or low mood.

Low ($T < 45$): sociable and extroverted; outgoing, gregarious, friendly and talkative; strong need to be around other people; mix well; intelligent, expressive, verbally fluent; active, energetic, vigorous; interested in status, power and recognition; seeks out competitive situations; have problem with impulse control; act without considering the consequences of actions; immature, self-indulgent; superficial, insincere relationships; manipulative, opportunistic; arouses resentment and hostility in others.

Adapted from Butcher (1989).

In addition, prior to publication of the MMPI-2 there were a number of validity studies conducted on the revised form. For example, the MMPI revision committee collected personality ratings on more than 800 couples included in the normative sample. These personality ratings clearly cross-validated a number of the original scales. Moreover, validation research was conducted on a number of samples including schizophrenics and depressives (Ben-Porath, Butcher, & Graham, 1991); marital problem families (Hjemboe & Butcher, 1991); potential child-abusing parents (Egeland, Erickson, Butcher, & Ben-Porath, 1991); alcoholics (Weed, Butcher, Ben-Porath, & McKenna, 1992); airline pilot applicants (Butcher, 1994); military personnel (Butcher, Jeffrey et al., 1990).

Since the MMPI-2 was published in 1989 a number of other validation studies have been published (Archer, Griffin, & Aiduk, 1995; Ben-Porath, McCully, & Almagor, 1993; Blake et al., 1992; Clark, 1996; Husband & Iguchi, 1995; Keller & Butcher, 1991; Khan, Welch, & Zillmer, 1993).

4.14.2.2 Basic Personality Inventory

The BPI (Jackson, 1989) was published as an alternative to the MMPI-2 for the global assessment of psychopathology. The key aims in developing the BPI were to produce a broad-band measure of psychological dysfunctioning as measured by the MMPI that was: (i) relatively short, (ii) incorporated modern principles of test construction, and (iii) showed

Table 3 Description of the MMPI-2 content scales.

1. Anxiety (ANX)
High scorers report general symptoms of anxiety including tension, somatic problems (i.e., heart pounding and shortness of breath), sleep difficulties, worries, and poor concentration. They fear losing their minds, find life a strain, and have difficulties making decisions. They appear to be readily aware of these symptoms and problems, are willing to admit to them.
2. Fears (FRS)
A high score indicates an individual with many specific fears. These specific fears can include blood; high places; money; animals such as snakes, mice, or spiders; leaving home; fire; storms and natural disasters; water; the dark; being indoors; and dirt.
3. Obsessiveness (OBS)
High scorers have tremendous difficulties making decisions and are likely to ruminate excessively about issues and problems, causing others to become impatient. Having to make changes distresses them, and they may report some compulsive behaviors such as counting or saving unimportant things. They are excessive worriers who frequently become overwhelmed by their own thoughts.
4. Depression (DEP)
High scorers on this scale show significant depression. They report feeling blue, uncertain about their future, and uninterested in their lives. They are likely to brood, be unhappy, cry easily, and feel hopeless and empty. They may report thoughts of suicide or wishes that they were dead. They may believe that they are condemned or have committed unpardonable sins. Other people may not be viewed as a source of support.
5. Health concerns (HEA)
Individuals with high scores report many physical symptoms across several body systems. Included are gastrointestinal symptoms (e.g., constipation, nausea and vomiting, stomach trouble), neurological problems (e.g., convulsions, dizzy and fainting spells, paralysis), sensory problems (e.g., poor hearing or eyesight), cardiovascular symptoms (e.g., heart or chest pains), skin problems, pain (e.g., headaches, neck pains), respiratory troubles (e.g., coughs, hay fever, or asthma). These individuals worry about their health and feel sicker than the average person.
6. Bizarre mentation (BIZ)
Psychotic thought processes characterize individuals high on the BIZ scale. They may report auditory, visual, or olfactory hallucinations and may recognize that their thoughts are strange and peculiar. Paranoid ideation (e.g., the belief that they are being plotted against or that someone is trying to poison them) may be reported as well. These individuals may feel that they have a special mission or powers.
7. Anger (ANG)
High scorers tend to have anger control problems. These individuals report being irritable, grouchy, impatient, hotheaded, annoyed, and stubborn. They sometimes feel like swearing or smashing things. They may lose self-control and report having been physically abusive towards people and objects.
8. Cynicism (CYN)
High scorers tend to show misanthropic beliefs. They expect hidden, negative motives behind the acts of others, for example, believing that most people are honest simply for fear of being caught. Other people are to be distrusted, for people use each other and are only friendly for selfish reasons. They likely hold negative attitudes about those close to them, including fellow workers, family, and friends.
9. Antisocial practices (ASP)
High scorers tend to show misanthropic attitudes like high scorers on the CYN scale. The high scorers on the ASP scale report problem behaviors during their school years and other antisocial practices like being in trouble with the law, stealing or shoplifting. They report sometimes enjoying the antics of criminals and believe that it is all right to get around the law, as long as it is not broken.
10. Type A (TPA)
High scorers report being hard-driving, fast-moving, and work-oriented individuals, who frequently become impatient, irritable, and annoyed. They do not like to wait or be interrupted. There is never enough time in a day for them to complete their tasks. They are direct and may be overbearing in their relationships with others.
11. Low self-esteem (LSE)
High scores on LSE characterize individuals with low opinions of themselves. They do not believe that they are liked by others or that they are important. They hold many negative attitudes about themselves including beliefs that they are unattractive, awkward, and clumsy, useless, and a burden to others. They certainly lack self-confidence, and find it hard to accept compliments from others. They may be overwhelmed by all the faults they see in themselves.

Table 3 (continued)**12. Social discomfort (SOD)**

High scorers tend to be very uneasy around others, preferring to be by themselves. When in social situations, they are likely to sit alone, rather than joining in the group. They see themselves as shy and dislike parties and other group events.

13. Family problems (FAM)

High scorers tend to show considerable family discord. Their families are described as lacking in love, quarrelsome, and unpleasant. They even may report hating members of their families. Their childhood may be portrayed as abusive, and marriages seen as unhappy and lacking in affection.

14. Work interference (WRK)

High scorers tend to show behaviors or attitudes that are likely to contribute to poor work performance. Some of the problems relate to low self-confidence, concentration difficulties, obsessiveness, tension and pressure, and decision-making problems. Others suggest lack of family support for the career choice, personal questioning of career choice, and negative attitudes towards co-workers.

15. Negative treatment indicators (TRT)

High scorers tend to show negative attitudes towards doctors and mental health treatment. High scorers do not believe that anyone can understand or help them. They have issues or problems that they are not comfortable discussing with anyone. They may not want to change anything in their lives, nor do they feel that change is possible. They prefer giving up, rather than facing a crisis or difficulty.

Adapted from: Butcher, Graham, Williams, & Ben-Porath (1990).

empirical evidence of being able to discriminate between normal and dysfunctional persons as well as being able to predict pathological behavior.

The BPI is made up of 240 items, grouped into 12×20 item scales. Neurotic tendencies are measured through the scales of Hypochondriasis, Depression, Anxiety, Social Introversion, and Self-depreciation. Aspects of sociopathy are measured by Denial, Interpersonal Problems, Alienation, and Impulse Expression scales. Psychotic behavior is assessed by scales labeled Persecutory Ideas, Thinking Disorder, and to some degree, by the Deviation scale. The Deviation scale comprises 20 critical items that are intended to serve as the basis for further clinical follow up. In contrast, the definitions of the constructs reflected in the other 11 items on the BPI are based on the results of a multivariate analysis of the content underlying the MMPI and the Differential Personality Inventory (Jackson & Messick, 1971).

The strong internal psychometric properties of the BPI attest to its careful construction (Jackson, 1989). Item properties and item factor analyses support the internal structure of the instrument. Both internal consistencies and test-retest reliability estimate fall in the 0.70 to 0.80 range. Various validity studies (some published in the manual, others in the literature) show that the BPI can indeed discriminate between normal and non-normal (e.g., delinquent) persons and can, within psychiatric populations, predict a variety of clinical criteria. Norms exist for adolescents and adults, spanning a variety of sample types such as community, psychiatric, college, and forensic.

However, these norms are almost entirely based on white populations. Moreover, most of the normative sample was collected using nonstandard data collection procedures. For example, booklets were mailed to subjects instead of being administered under standard conditions. In addition, each of the subjects in the normative sample responded to one-third of the items in the booklet, an artifact that makes it difficult to perform some analyses on the normative sample (e.g., alpha coefficients).

Among the difficulties that have been associated with the BPI is a lack of validity scales for identifying invalid response protocols. The one content scale that would logically appear to have some bearing on the issue of protocol validity is the Denial scale, which is described in the manual as a measure of lack of insight and lack of normal affect. Unfortunately, Denial appears to be a relatively weak scale in terms of its reliability and validity in various empirical studies (Holden et al., 1988; Jackson, 1989). Two other complaints that have been voiced are a lack of an established link to diagnostic categories and a lack of work on profile interpretation.

Overall, the BPI has shown psychometric potential as a general measure of psychopathology. The basic developmental work on it is sound. What the BPI now needs are more extensive norms along with further work on its clinical applicability.

4.14.2.3 Personality Assessment Inventory

The PAI (Morey, 1991) is another inventory of general psychopathology. The PAI is a 344

item self-report measure designed to screen for approximately the same pathological domains as the MMPI/MMPI-2. It is used to collect information related to diagnosis, and to provide input on treatment planning.

The PAI includes four validity scales, 11 clinical scales, five treatment scales and two interpersonal scales. The clinical scales are Somatic Complaints, Anxiety-related Disorders, Depression, Mania, Paranoia, Schizophrenia, Borderline Features, Antisocial Features, Alcohol Problems, and Drug Problems. Treatment scales are Aggression, Suicidal Ideation, Stress, Nonsupport, and Treatment Rejection. Interpersonal scales are Dominance and Warmth. Of these scales, 10 are further divided into subscales that are intended to measure distinct constructs. A total of 27 items are designated critical items which, according to the author, should be followed up. This does seem to be a relatively large number of scales to interpret meaningfully in view of the total number of items.

Evidence for the reliability of the 22 scales is generally good. Across normative, clinical and college samples, median alpha coefficients are all in the 0.80–0.90 range. One month test–retest coefficients are reported to be in the 0.80s in the manual and in the 0.70s in the literature (Boyle & Lennon, 1994). Interestingly, some of the lowest reliabilities are found for the validity scales, a phenomenon that has been found for the validity scales on other instruments such as the MMPI (e.g., Fekken & Holden, 1991) and may well be a function of range restriction in the scores on such scales.

Normative PAI data for the USA are extensive. Standardization samples for a census-matched group, a clinical sample representing 69 sites, and a college sample drawn from seven different US universities each number over 1000 respondents. Normative data in the manual are also reported separately by age, education, gender, and race. Normative information for other countries (Canada, Australia, the UK, etc.) would be desirable.

In view of the recency of the publication of the PAI, there are only a handful of studies that bear on its validity. The manual reports evidence of the concurrent validity of the PAI in the form of correlations with other measures of psychopathology. A number of other studies show that the PAI can discriminate between diagnostic groups. Boyle and Lennon (1994) showed that the PAI can distinguish normals, alcoholics, and schizophrenics. Schinka (1995) was further able to use the PAI to develop a typology for alcoholics that had validity with several external variables. Finally, Alterman and colleagues (1995) demonstrated that metha-

done maintenance patients scored differently from the normative populations.

Generally, preliminary data suggest that the PAI is a well-constructed and brief measure of psychopathology. More research on its clinical validity needs to be completed before it can be considered to be a useful measure of psychopathology in clinical settings. There are no data to support the PAI's use in lieu of the MMPI-2 which has a more substantial empirical database.

4.14.3 SPECIALIZED OR FOCUSED CLINICAL ASSESSMENT MEASURES

In this section we will address several other measures that have been developed for clinical assessment and research to assess specific or more narrowly focused characteristics rather than omnibus instruments such as the MMPI-2. Several of these instruments will be examined to illustrate their application and potential utility for evaluating specific problems in clinical settings.

4.14.3.1 Millon Clinical Multiaxial Inventory

The MCMI (Millon, 1977, 1987, 1994) was developed by Theodore Millon for making clinical diagnoses on patients. The MCMI was intended to improve upon the long-established MMPI. In contrast to the MMPI/MMPI-2, the MCMI was designed with fewer items; is based on an elaborate theory of personality and psychopathology; and explicitly focuses on diagnostic links to criteria from the *Diagnostic and statistical manual of mental disorders (DSM)*.

The MCMI was developed rationally rather than empirically. Millon has stated in his three test manuals (1977, 1987, 1994) as well as elsewhere (Millon & Davis, 1995) that development of the MCMI is to be an ongoing process. To keep the MCMI maximally useful for clinical diagnosis and interpretation, it must be continually updated in view of theoretical refinements, empirical validation studies, and evolutions in the official *DSM* classification systems. Most updated test manuals leave the test user with the impression that the developer considered test revision a necessary evil. Very rarely do you see continuous improvement as a test developer's goal, in part because this ever-changing process makes the accumulation of a solid research base difficult.

All three MCMI versions comprise 175 true/false items. However, across versions, the exact test items have evolved through revision or replacement. The number of scales and validity

indices that can be calculated from these items has also increased. The original version, the MCMI-I, had 20 clinical scales and two validity scales. The MCMI-II yielded 22 clinical scales and three validity scales. The current MCMI-III has 24 clinical scales, three modifying indices and a validity index. Many items appear on several scales making for great item overlap.

On the MCMI-III, 14 clinical scales assess personality patterns that relate to *DSM-IV* Axis II disorders. Another 10 scales measure clinical syndromes related to *DSM-IV* Axis I disorders. The modifying indices, Disclosure, Desirability, and Debasement, are correction factors applied to clinical scale scores to ameliorate respondents' tendencies to distort their responses. The validity index comprises four bizarre or highly improbable items meant to detect careless, random, or confused responding.

The relationship between scales and items is explicated in detail in the manual. Millon started with a theory-based approach to writing items, followed by an evaluation of the internal structure of the items, and finally engaged in an assessment of the diagnostic efficiency of each item for distinguishing among diagnostic groups before final placement of an item on a scale. Millon departed from usual psychometric practice in a way that results in some unfortunate complications. Item overlap across scales is permitted: on average, items appear on three different scales with differential weights. This makes scoring inordinately complex, which makes assessment of scale homogeneity complex, and in turn this makes evaluation of the empirical structure underlying the scales complex. There are technical solutions for these problems but the result is that the MCMI is not an easy instrument with which to work.

Despite its psychometric drawbacks, the theory underlying the MCMI is generally agreed to be elegant and a substantial asset (McCabe, 1987; Reynolds, 1992). Each dimension measured by the test has a clear conceptual link to Millon's theory of psychopathology. Such a theory allows for the generation of clinical inferences based on a small number of fundamental principles (Millon & Davis, 1995). Not only do these inferences guide measurement, but also they enhance understanding of the constructs, bear on practical treatment decision, and produce research hypotheses.

One of the stated goals of the MCMI is to place patients into target diagnostic groups. To this end, the MCMI scales are directly coordinated with the *DSM* diagnostic categories. How well does the MCMI live up to its aim? The manuals report good evidence of diagnostic efficiency. However, the recent literature (available on the MCMI-I and MCMI-II) suggests

the following three generalizations. First, the MCMI has only modest accuracy for assigning patients to diagnostic groups across a variety of clinical criteria (e.g., Chick, Martin, Nevels, & Cotton, 1994; Chick, Sheaffer, Goggin, & Sison, 1993; Flynn, 1995; Hills, 1995; Inch & Crossley, 1993; Patrick, 1993; Soldz, Budman, Demby, & Merry, 1993). Second, the MCMI may be better at predicting the absence than the presence of a disorder (Chick et al., 1993; Hills, 1995; Soldz et al., 1993). Third, the MCMI may be better at predicting some types of disorders than others but there is little agreement on which ones (Inch & Crossley, 1993; Soldz et al., 1993).

One source of the difficulty may be the base rate scores. Raw scores on scales are weighted and converted to base rate scores. The base rate scores reflect the prevalence of a particular personality disorder or pathological characteristic in the overall population. Their use is intended to maximize the number of correct classifications relative to the number of incorrect classifications when using the MCMI to make diagnoses (Millon & Davis, 1995). If the estimated base rates for the various diagnostic categories are poor, then the predictive accuracy of the MCMI can be expected to be poor (Reynolds, 1992).

One negative consequence of the type of norms used in the development of the MCMI inventories is that they do not discriminate between patients and normals. Use of the MCMI assumes that the subject is a psychiatric patient. Consequently, the MCMI overpathologizes individuals who are not actually patients. The MCMI should not be used where issues of normality need to be addressed. For example, if the test were used in family custody evaluations or personnel screening, the test interpretation would appear very pathological—it cannot do otherwise.

How does the MCMI compare to the MMPI? The MCMI publisher appears to emphasize that the MCMI and MMPI-2 measure different characteristics and the MCMI is shorter to administer to patients. Whereas the MMPI measures a broad range of psychopathology, the MCMI has its premier focus on the assessment of personality disorders. Consonant with its rational construction, the elaborate theoretical underpinnings of the MCMI are impressive. In contrast, however, the test literature that supports the validity of the MMPI/MMPI-2 is not available for the MCMI. Validation research on it has not proceeded at a very high pace. Whether the MCMI will have either the clinical utility or the heuristic value that the MMPI enjoys remains unanswered until more clinical research, and perhaps more refinements, are undertaken with the MCMI.

4.14.3.2 The Beck Depression Inventory

The BDI was first introduced in 1961, and it has been revised several times since (Beck et al., 1988). The BDI has been widely used as an assessment instrument in gauging the intensity of depression in patients who meet clinical diagnostic criteria for depressive syndromes. However, the BDI has also found a place in research with normal populations, where the focus of use has been on detecting depression or depressive ideation.

The BDI was developed in a manner similar to the MMPI: clinical observations of symptoms and attitudes among depressed patients were contrasted to those among nondepressed patients in order to obtain differentiation of the depressed group from the rest of the psychiatric patients. The 21 symptoms and attitudes contained in the BDI reflect the intensity of the depression; items receive a rating of zero to three to reflect their intensity and are summed linearly to create a score which ranges from 0 to 63. The 21 items included reflect a variety of symptoms and attitudes commonly found among clinically depressed individuals (e.g., Mood, Self-dislike, Social Withdrawal, Sleep Disturbance). The BDI administration is straightforward, and it can be given as an interview by the clinician or as a self-report instrument (requiring a fifth or sixth grade reading level).

The BDI is interpreted through the use of cut-off scores. Cut-off scores may be derived based on the use of the instrument (i.e., if a clinician wishes to identify very severe depression, then the cut-off score would be set high). According to Beck et al. (1988), the Center for Cognitive Therapy has set the following guidelines for BDI cut-off scores to be used with affective disorder patients: scores from 0 through 9 indicate no or minimal depression; scores from 10 through 18 indicate mild to moderate depression; scores from 19 through 29 indicate moderate to severe depression; and scores from 30 through 63 indicate severe depression.

Two important issues must be considered by clinicians regarding the results of the BDI. Unlike the MMPI/MMPI-2 and other major self-report instruments, the BDI has no safeguards against faking, lying, or variable response sets. Thus, clinicians are warned against this drawback of the BDI in assessing depressive thoughts and symptoms. In settings where faking or defensiveness are probable threats to the validity of the test, clinicians may need to reconsider their use of the BDI. The other issue pertains to the state–trait debate in assessment. The BDI is extremely sensitive to differences in the instructions given to an examinee such that

certain instructions yield a state-like index of depressive thinking, whereas, other instructions yield a more trait-like index of depressive thinking. Again, clinicians are encouraged to use caution when administering the BDI and to tailor the administration instructions to the type of index (state or trait) that is desired.

In their review of the psychometric properties of the BDI, Beck et al. (1988) reported high internal consistency reliability of the instrument among both psychiatric and nonpsychiatric populations. The authors also reported that the BDI closely parallels the changes in both patient self-report and clinicians' ratings of depression (i.e., the BDI score accurately reflects changes in depressive thinking). Finally, they also presented evidence for the content, concurrent, discriminant, construct, and factorial validity of the BDI.

The acceptable reliability and validity of the BDI have helped make it a widely used objective index of depressive thinking among clinicians. Perhaps the most obvious use of the BDI is as an index of change in the level or intensity of depression. With an increasing focus on managed healthcare and accountability by psychotherapeutic service providers, the BDI offers a reliable and valid index of depressive symptoms and attitudes which can be used effectively to document changes brought about in therapy.

4.14.3.3 The State-Trait Anxiety Inventory

The STAI was developed by Spielberger et al. (1970) to measure anxiety from the perspective of states vs. traits. The state measurement assesses how the individual feels “right now” or at this moment. Subjects are asked to rate the intensity of their anxious feelings on a four point scale as to their experience of feelings in terms of: not at all, somewhat, moderately so, or very much so. The trait anxiety measure addresses how the individuals generally feel by rating themselves on a four-point scale: almost never, sometimes, often, or almost always.

Since it was developed in 1966 the STAI has been translated into over 48 different languages and has been widely researched in a variety of clinical and school settings (Spielberger, Ritterband, Sydeman, Reheiser, & Unger, 1995). The evidence for construct validity of the STAI comes from a variety of sources, for example, correlations with other anxiety measures (Spielberger, 1977), clinical settings (Spielberger, 1983), and medical and surgical patients (Spielberger, 1976). Nonetheless, there has been general debate in the literature about the conceptual and practical benefits of the trait vs. state distinction.

The STAI has become a very widely used measure in personality and psychopathology research in the USA and in other countries. However, the state-trait inventories have not been as broadly used as clinical assessment instruments.

4.14.3.4 Whitaker Index of Schizophrenic Thinking

The WIST (Whitaker, 1973, 1980) was developed to measure the type of thought impairment that differentiates between schizophrenic and "normal" thought processes. The WIST is intended to be individually administered as a screening tool or as one part of a battery of tests. Its multiple choice format makes the WIST a test that is easy to give and to score.

On the WIST, schizophrenic thought is defined as a discrepancy between actual and potential performance on cognitive reasoning tasks. This impaired thinking has three components: (i) a degree of illogicality, as reflected in inappropriate associations and false premises; (ii) a degree of impairment relative to previous performance, as reflected in slowness; and (iii) a degree of unwittingness, as reflected in unawareness of the incorrectness of responses. Whitaker's definition of schizophrenic thought is carefully explicated in the manual, making it possible for the test user to understand exactly what the WIST is measuring. Whitaker, however, has been criticized for basing his definition on a narrow reading of the literature on schizophrenic thought disorder (Payne, 1978).

The WIST itself is made up of 25 multiple choice items that are divided into three subtests: Similarities, Word Pairs, and New Inventions. Each item consists of a stimulus and five response options that differ in degree of illogicality. To illustrate, consider this sample item for the Similarities subtest: car: automobile, tires, my transportation, jar, smickle. The correct answer receives a score of 0; a loose association, reference idea, clang association, and nonsense association would receive scores of 1, 2, 3, and 4, respectively. The test administrator presents for a second time any items that the respondent answered incorrectly on the initial test taking.

Three scores are calculated for the WIST: total Score for all response alternatives selected either in the original or in the second enquiry phase; total Time for the initial completion of the WIST; and an overall Index that combines WIST Score and WIST Time. Presumably the WIST Score and Time components are added together because they both relate to schizophrenic thought disorder. However, a review of the empirical data supporting this

conceptualization finds mixed results at best (Fekken, 1985). One view is that the predictive efficiency of the WIST is not enhanced by including the Time component.

There are two parallel forms of the WIST which differ in content. Items on Form A are intended to be stressful or anxiety provoking. They assess oral-dependency, hostility, and manifestly sexual content. The content of Form B is neutral. There has been little formal evaluation of the comparability of the two forms. Although some studies have reported similar rates of diagnostic efficiency for the two forms (Evans & Dinning, 1980; Leslie, Landmark, & Whitaker, 1984), overall validity evidence would suggest that Form A is stronger than Form B. Additional work to clarify the comparability of Forms A and B would have implications for selecting which form to use and for using these alternate WIST forms to assess changes in symptomatology.

There is relatively little information on reliability available on the WIST. The manual does report intratest reliabilities of the two WIST forms as Hoyt's reliability coefficients of around 0.80. Test-retest reliability data, either on the WIST subscores or subscales, are not provided in the manual nor do they appear to be readily available in the literature. Similarly, alternate forms of reliability, calculated as correlations between subscales, or subscores, are not readily available.

There is reasonable evidence for the convergent validity of the WIST. Two reviews report that WIST scores tend to have a 60–70% agreement with systems for diagnosing schizophrenia (Fekken, 1985; Grigoriadis, 1993). The WIST has been empirically associated with other indices of schizophrenia such as the MMPI/MMPI-2 Sc scale (Evans & Dinning, 1980; Fishkin, Lovallo, & Pishkin, 1977; Grigoriadis, 1993) although not with other schizophrenia indices including conceptually relevant SCL-90 scales (Dinning & Evans, 1977) and the New Haven Schizophrenic Index (Knight, Epstein, & Zielony, 1980).

Discriminant validity of the WIST has been harder to establish. The WIST has difficulty distinguishing among different psychiatric diagnostic groups (e.g., Burch, 1995; Pishkin, Lovallo, & Bourne, 1986). This is a serious shortcoming because the WIST is likely to be used in clinical settings precisely to help make such distinctions. A second problem with discriminant validity has been the tendency of the WIST to correlate negatively with measures of general cognitive ability. Based on such data one could share the view of at least one pessimistic reviewer and claim that the WIST has no demonstrated use (Payne, 1978).

Alternatively, the WIST may have a role in a comprehensive assessment battery. The WIST has never been promoted as a stand-alone test, nor has it ever been promoted as a comprehensive measure of the full range of schizophrenic symptomatology. Rather, the data on the WIST may be thought of as a general measure of cognitive deficit rather than cognitive deficit specific to schizophrenia. Thus, it may provide one objective source of data for accepting or rejecting a more general diagnosis of psychosis.

4.14.4 NORMAL RANGE PERSONALITY ASSESSMENT

Identification and description of psychological disorder is but one reason to administer a personality measure to an adult. Objective personality measures also are widely used outside the clinic or hospital setting for normal range assessment situations. Several objective personality measures will be discussed regarding their use in research, educational, vocational assessment, and personnel selection.

4.14.4.1 Objective Personality Measures in Research

4.14.4.1.1 *The Five Factor Model (the Big Five)*

One of the most popular current conceptualizations being studied today involves taxonomies of personality traits based on factor analytic methods and is commonly referred to as the "Big Five," or the Five Factor Model (FFM). As Butcher and Rouse (1996) note in their review, some personality researchers have rejected the FFM as an end-all to personality trait theory, while others in the area of personality research continue to embrace it. Among the proponents of the FFM are Costa and his colleagues, who have proposed the NEO Personality Inventory (NEO-PI) as a self-report measure of the FFM personality dimensions (Costa & McCrae, 1992). The FFM consists of five factor-analytically derived dimensions of personality. The five dimensions (are Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). Neuroticism has long been a familiar adjective among clinicians for describing people who tend to experience psychological distress. At the opposite end of the Neuroticism dimension is Emotional Stability, which represents the tendency to stay on a psychologically even keel. Extroversion encompasses the concepts of positive emotionality, sociability, and activity. The Extroversion dimension is rather broad in its scope, although

lay people often use the term narrowly to refer to general "outgoingness." Openness to Experience represents, broadly, a person's level of constriction in their experiencing of the world; it is often associated with creativity (and even hypnotizability). Agreeableness represents the dimension of interpersonal behavior; the counterpart of Agreeableness is Antagonism. Finally, Conscientiousness represents a dimension of scrupulous organization of behavior.

4.14.4.1.2 *NEO Personality Inventory*

The NEO-PI is a 181-item inventory designed to index the FFM personality dimensions (N, E, O, A, and C); the NEO-PI also yields several subscales of the N, E, and O dimensions (Costa & McCrae, 1992). There is a self-report form of the NEO-PI as well as an observer-rating form. Although the authors of the NEO-PI argue for its utility in clinical settings, the instrument has been studied almost exclusively in nonclinical populations. One example of a recent investigation using the NEO-PI in research with psychiatric samples is the examination of differences in stimulus intensity modulation among older depressed individuals with and without mania features (Allard & Mishara, 1995). The NEO-PI was used to examine the hypothesis that stimulus intensity augmenters with unipolar depression would be introverted, whereas reducers with bipolar depression would be extroverted. Determinations of depression with and without mania features were based on scores on the MMPI. Costa and McCrae (1992) point to the NEO-PI as a useful assessment tool in aiding the clinician with understanding the client, selection of treatment, and even anticipating the course of therapy. However, the NEO-PI has not found wide use among clinicians to date (Butcher & Rouse, 1996).

Unfortunately, the NEO-PI is very susceptible to faking (Bailey & Ross, 1996) and does not contain validity indices to detect deviant response sets.

4.14.4.1.3 *Multidimensional Personality Questionnaire*

The Multidimensional Personality Questionnaire (MPQ) is a 300-item self-report instrument that was developed by Tellegen (unpublished manuscript) in an attempt to clarify the "self-view domain" in personality research. Tellegen used a classical iterative test construction approach involving several rounds of factor analysis to come up with the 11 primary scales, six validity scales, and three "higher-order"

factors. The 11 primary scales (which include the dimensions of Social Potency, Control, Harm Avoidance, Well-being, Aggression, and others) load onto the three higher-order scales, which represent the familiar personality domains of Positive Affectivity, Negative Affectivity, and Constraint. The six validity indices include VRIN and TRIN, which are conceptually similar to those on the MMPI-2.

The MPQ is not as widely used as other objective personality measures in either the clinical or the research domain. However, recent investigations suggest a place for the MPQ in normal range and clinical personality assessment. For example, Kuhne, Orr, and Baraga (1993) demonstrated the utility of certain MPQ scales for discriminating among veterans with and without post-traumatic stress disorder. Krueger et al. (1994) administered the MPQ to adolescents in their community-based longitudinal study and found that certain MPQ scales were useful in distinguishing those who engaged in delinquency from those who abstained. These two reports suggest the utility of the MPQ both in clinical settings and in normal range personality assessment.

4.14.4.1.4 Personality Research Form

One well-known measure of normal personality is the Personality Research Form (PRF). Developed by Douglas N. Jackson (1984), the PRF is a true/false, multiscale measure of 20 of the psychosocial needs (e.g., achievement, aggression, sociability) originally defined by Henry Murray (1938). Many psychometrics texts hold up the PRF as a model of the construct approach to test construction. Indeed, much of the appeal of the PRF lies in its ability to measure a large number of normal personality characteristics while minimizing both scale intercorrelations and the influence of social desirability and acquiescence. Moreover, a variety of validity studies have been published in the 1980s and 1990s supporting the psychometric soundness of the PRF. Critics of the PRF complain that despite its technical elegance the PRF does not reflect an integrated model of personality, which limits its applicability to real-life testing situations. Recent research, however, shows that the content assessed by the PRF may be well described by the Five Factor structure (Paunonen, Jackson, Trzebinski, & Fosterling, 1992). In research the popularity of the PRF remains high as attested to by the number of references in bibliographies, such as the one produced by MacLennan in 1991, which lists over 375 studies featuring the PRF in the literature.

4.14.4.1.5 Sixteen Personality Factor Test

The 16PF was originally developed in the 1940s by Raymond Cattell to measure the primary factors of normal personality. At that time, Cattell's unique contribution was to apply factor analysis as a method for uncovering the full scope of personality. The fifth and current edition of the 16PF (Cattell, Cattell, & Cattell, 1993) measures the well-known 16 personality factors, plus it summarizes these factors into five global factors which again bear similarity to the well-known "Big Five." The global factors are: extroversion, anxiety, tough-mindedness, independence, and self-control. Relative to earlier editions, the fifth editions includes updated language; fewer items; and improved reliability and response style scales to measure impression management, infrequent responding, and acquiescence. Because of the recency of the publication of the fifth edition, there are few studies pertaining to its validity. However, a large database supports the validity of earlier editions of the 16PF. The 16PF has applicability in clinical, educational, organizational, and research settings. With the expansion of its interpretive reports and profiles, the 16PF would appear to be particularly useful in personal or vocational counseling settings.

4.14.4.1.6 California Psychological Inventory

The CPI developed by Gough (1957) is a multiscale, objective, self-report instrument used with normal range and psychiatric populations (Megargee, 1972). The CPI is similar in structure and content to the MMPI (and the MMPI-2). In fact, many of the items on the CPI are identical in wording to the items on the MMPI. The CPI focuses on "everyday" concepts about personality, such as dominance and responsibility among others. Eighteen scales (divided into four classes) are derived from the CPI; the results of the test are interpreted by reference to the plotted profile of standard scores. The body of literature using the CPI in normal range assessment is quite extensive in both quantity and breadth. Investigations include the use of the CPI to identify personality types, based on profiles, among a group of college students (Burger and Cross, 1979) and an examination of the underlying personality structure as measured by the CPI, again using a sample of college students (Deniston and Ramanaiah, 1993). The CPI also has been employed in the assessment of psychiatric samples. Especially noteworthy is the utility of the CPI with criminal samples (see Laufer, Skoog, and Day, 1982, for a relevant review of this literature).

Some specific applications of psychological tests in “normal range” settings are now described.

4.14.4.2 Objective Personality Measures in Educational/Vocational Assessment

The use of objective personality measures has become increasingly popular among professionals in the field of educational/vocational assessment. Research has identified the FFM, the NEO-PI, and the CPI as particularly useful in educational/vocational assessment.

4.14.4.2.1 FFM

The Big Five personality dimensions of the FFM have been studied in several contexts related to educational/vocational assessment. Moreover, because the FFM is a theoretical concept, various instruments have been used to assess the relationship between the Big Five personality dimensions and various educational/vocational assessment variables. Two recent investigations illustrate the utility of the Big Five personality dimensions in identifying candidates for admission to educational institutions and identifying characteristics among students in various university programs. Williams, Munick, Saiz, and Formy-Duval (1995) found that a mock graduate school admissions board (composed of graduate school faculty) favored for admission those hypothetical candidates whose applications reflected high ratings on the Big Five dimensions of Conscientiousness and Openness to Experience. Conversely, the Big Five dimensions of Extroversion and Agreeableness were not associated with a favorable impression of hypothetical candidates. Kline and Lapham (1992) assessed the Big Five personality dimensions among a group of college students to examine personality differences among the various fields of study (i.e., between different college majors). Students of various majors were not discriminated by levels of either Neuroticism or Extroversion. However, students in two fields of study (science and engineering) were marked by high ratings on the Big Five personality dimensions of Conscientiousness and Conventionality.

4.14.4.2.2 NEO-PI

The personality dimensions that underlie the NEO-PI appear to be related to at least one well-known typology of vocational personalities (which, in turn, correspond to vocational preferences). Gottfredson, Jones, and Holland (1993) found a relationship between the Big Five

personality dimensions assessed by the NEO-PI and the six vocational personality dimensions proposed by Holland (assessed with the Vocational Preferences Inventory). In general, there was overlap among two to four of the significant factors extracted from each of the assessment instruments. However, the NEO-PI Neuroticism, Likability, and Control factors were not represented in the Holland vocational personality dimensions, which suggests a distinctive and qualitatively different role for objective personality assessment in vocational counseling. Other work with the NEO-PI has shown that two of the Big Five personality dimensions, Neuroticism and Agreeableness, are strongly related to occupational burnout among health-care workers (Piedmont, 1993). Healthcare workers with higher ratings on the Neuroticism dimension were more likely to experience occupational burnout. Conversely, workers with higher ratings on the Agreeableness dimension were less likely to succumb to occupational burnout.

4.14.4.2.3 CPI

The CPI is one of the most widely used personality instruments in normal range assessment in educational/vocational settings. Walsh (1974) examined personality traits among college students identified as making hypothetical career choices that were either congruent or incongruent with their vocational personality type (as proposed by Holland). It was found that congruent students could be described by their CPI profiles as socially accepted, confident, and planful, whereas incongruent students could be described as impulsive, unambitious, and insecure. The well-known Strong Vocational Interest Blank (SVIB) used commonly among vocational and educational counselors is related to various personality traits as well. Johnson, Flammer and Nelson (1975) found a relationship among SVIB factors and CPI personality factors, especially those related to the global introversion/extroversion personality dimension.

4.14.4.2.4 MMPI/MMPI-2

Although it was not designed for educational research or placement purposes the MMPI/MMPI-2 has been among the most frequently employed instruments in this context. In a recent survey of test use in personnel and educational screening the MMPI has been employed effectively in a number of studies, for example: Anderson (1949), Appleby and Haner (1956), Applezweig (1953), Barger and Hall (1964), Barthol and Kirk (1956), Burgess

(1956), Centi (1962), Clark (1953, 1964), and Frick (1955) to mention only a few.

It appears evident that personality assessment contributes information independent of direct educational/vocational interest measures. There is extensive information available on the use of the MMPI/MMPI-2 in this setting. As demonstrated in the research with the NEO-PI and the CPI, objective personality assessment instruments yield information relevant to vocational assessment over and above that given by narrowly defined vocational personality assessments (like the one proposed by Holland). Moreover, elements of the FFM and independent personality constructs of the CPI have proven their utility in educational assessment in diverse areas such as admissions to educational institutions and choice of major fields of study. Clearly, objective personality assessment has carved out a valuable niche in normal range educational/vocational assessment.

4.14.4.3 Personnel Screening

Among the earliest and most extensive uses of personality tests with normals has been for the purpose of employment screening. The first formal North American, English language personality inventory, the Woodworth Personnel Data Sheet, was developed to screen out unfit draftees during World War I. A number of other personality questionnaires were developed in the 1930s to aid in personnel selection decisions.

The development and use of the MMPI during World War II provided a means for assessment psychologists to detect psychological problems that might make people unsuitable for key military assignments. Early research on the use of the MMPI centered around pilot selection and selection of nuclear submarine crewman. Following World War II, the MMPI came to be widely used in personnel selection particularly for occupations that required great responsibility or involved high stress such as air flight crews (Butcher, 1994; Cerf, 1947; Fulkerson, Freud, & Raynor, 1958; Fulkerson & Sells, 1958; Garetz & Tierney, 1962; Geist & Boyd, 1980; Goorney, 1970; Jennings, 1948); police and other law enforcement personnel (Beutler, Nussbaum, & Meredith, 1988; Beutler, Storm, Kirkish, Scogin, & Gaines, 1985; Butcher, 1991; Dyer, Sajwaj, & Ford, 1993; Hargrave & Hiatt, 1987; Saxe & Reiser, 1976; Scogin, & Beutler, 1986; Scogin & Reiser, 1976); and nuclear power employees (Lavin, Chardos, Ford & McGee, 1987).

Following the revision of the MMPI and publication of the MMPI-2, the revised instrument has been used in personnel screening

(Butcher & Rouse, 1996) and currently is the most frequently used personality measure in personnel screening situations, particularly when the position is one that requires good mental health, emotional stability, and responsible behavior.

The MMPI-2 is usually employed in personnel selection to screen out candidates who are likely to have psychological problems from critical occupations such as police officer, airline pilot, and nuclear power control rooms, fire department, or air traffic control personnel.

4.14.4.4 Other Personality Measures in Personnel Selection

Objective measures of personality characteristics have a role in personnel selection similar to that in educational/vocational assessment. Namely, professionals in the field of personnel selection are interested in knowing which personality variables aid in the selection of quality employees. The literature suggests a valuable role for objective personality measures in the normal range assessment field of personnel selection.

4.14.4.5 The 16PF

One of the most widely used personality scales for employment screening is the 16PF. This inventory, with broad-ranging employment-relevant personality items, has been widely used in different contexts including: law enforcement (Burbeck & Furnham, 1985; Fabricatore, Azen, Schoentgen, & Snibbe, 1978; Hartman, 1987; Lawrence, 1984; Lorr & Strack, 1994; Topp & Kardash, 1986); pilots (Cooper & Green, 1976; Lardent, 1991); cabin crew personnel (Furnham, 1991); managers (Bartram, 1992; Bush, & Lucas, 1988; Chakrabarti, & Kundu, 1984; Henney, 1975); occupational therapists (Bailey, 1988); church counselors (Cerling, 1983) and teachers (Ferris, Bergin, & Wayne, 1988). The 16PF provides information about personality functioning and is typically employed to screen employees for positive personality features.

4.14.4.6 FFM

Barrick and Mount (1991) conducted a meta-analysis of the Big Five personality dimensions and their relationship to three criterion variables (job proficiency, training proficiency, and personnel data) within five occupational groups (professionals, police, managers, skilled/semiskilled and sales). Conscientiousness was related to each of the five occupational

groups. Additionally, Conscientiousness was related to the three criterion variables. These findings lead the authors to conclude that Conscientiousness is a personality trait related to job performance across occupational types. In a similar vein, Dunn, Mount, Barrick, and Ones (1995) found that the Conscientiousness personality dimension was related to managers' ratings of applicant hireability. This held true for various job types (medical technologist, carpenter, secretary, etc.). Taken together, these two reports suggest that professionals in the field of personnel selection can gain valuable information from objective personality measures of the FFM. Specifically, the Conscientiousness dimension, when taken with other relevant application information, may be a valuable discriminator among prospective job applicants.

4.14.4.7 NEO-PI

While research suggests that the Big Five personality dimensions add qualitative information to the personnel selection process, it may be that the various instruments used to measure the Big Five have differential validity across populations. Schmit and Ryan (1993) administered a shortened version of the NEO-PI to a sample of college students and to a sample of government job applicants. The FFM structure fitted the student population better than it fitted the job applicant population, which suggests a note of caution in putting too much weight on the Big Five personality dimensions in personnel selection. The authors note that job applicants are under different situational demands, which may affect their approach to a personality questionnaire (e.g., they may adopt a defensive response style). The lack of validity scales to assess response styles clearly limits the NEO-PI for this application. A recent study by Bailey and Ross (1996) showed that the NEO-PI is quite vulnerable to faking and is limited in not having scales to detect deviant response attitudes.

4.14.4.7.1 CPI

The NEO-PI and other personality measures have been examined in personnel selection studies covering a wide range of occupational groups (managers, secretaries, carpenters, etc.). However, the CPI has a long tradition of use in the selection of a specific occupation: law enforcement officers. Hargrave and Hiatt (1989) examined the ability of the CPI scales to differentiate police cadets rated by their instructors as suitable or unsuitable for the job of law enforcement officer. Several scales

(including Sense of Well-being, Sociability, and Social Presence) differentiated suitable from unsuitable cadets. In the second study of the report, the authors compared incumbent law enforcement officers who either had or had not experienced serious on-the-job problems (e.g., providing drugs to inmates, excessive use of force). The Socialization scale was among the best discriminators between groups. The authors concluded that the CPI is a useful aid to the selection of law enforcement officers.

In addition to the literature on law enforcement screening, the CPI may be useful as an index of work performance in other fields. Toward this end, Hoffman and Davis (1995) validated the Work Orientation and Managerial Potential scales for the CPI on groups of employees in an entertainment facility. However, several of the original CPI scales performed as well as the two new scales in predicting job performance, which questions the need for additional CPI scales in the selection of personnel.

In summary, the use of objective personality measures in the selection of job applicants has proven a worthwhile endeavor overall. The Big Five personality dimensions, especially Conscientiousness, might be useful to the personnel selection process especially in the assessment of conscientiousness. Professionals involved in the screening of law enforcement candidates can use the CPI to differentiate suitable from unsuitable candidates. The CPI may also be useful in discriminating among groups of job applicants (although its primary use has been in the law enforcement field).

The scales on the 16PF have been shown to be relevant to personality descriptions that are useful in personnel selection. Finally, when it comes to evaluating potential psychopathology in potential employees the MMPI-2 is usually the instrument most employed.

Although objective personality measures appear to have a place in personnel selection, all information on the candidate must be weighed, as several personality researchers have noted the problem of response sets associated with the situational demands of the application process. Specifically, job applicants may feel under pressure to make a good impression on their prospective employer and they may subsequently present themselves as overly virtuous or defensive.

4.14.5 SUMMARY

Human fascination with the concept of personality lead to the nineteenth century invention of the first objective personality test.

As with those very early personality measures, many of today's objective personality inventories are self-reports. The clinician's first concern when utilizing an objective personality measure is whether or not a client can accurately reveal information about his or her personality through a self-report instrument. Many objective personality instruments incorporate indices of test-taking attitudes (e.g., the Lie scale of the MMPI/MMPI-2) which allow the clinician to gauge a client's level of insight and willingness to self-disclose. Additionally, research shows that clients who do cooperate with testing produce personality profiles that match external criteria (e.g., clinician's notes and observations regarding the patient). Essentially, most clients are able to reveal their personalities competently through self-report measures.

Once a client is able to self-disclose information regarding his or her personality, the scale scores that are produced appear to be quite stable over time. Thus, most of the objective personality measures manage to capture trait (as opposed to state) characteristics. Five factors exhibit influence on the stability of personality: (i) instrument characteristics, (ii) length of retest interval, (iii) operationalization of personality as a stable construct for test construction, (iv) the extent to which a particular personality construct is associated with stability, and (v) person variables of the test-taker.

Several objective personality measures are designed to assess adults in clinical settings. The most well-known and widely used objective personality inventory in clinical settings is the MMPI-2, which provides a comprehensive survey of personality characteristics and clinical problems. The Basic Personality Inventory (BPI) and the Personality Assessment Inventory (PAI) are alternatives to the MMPI-2, but neither is as widely used as the MMPI-2. Several other objective measures have been developed for specialized or focused clinical use including the Millon Clinical Multiaxial Inventory (MCMI) designed for making personality diagnoses; the Beck Depression Inventory (BDI), which assesses depressive ideation; the State-Trait Anxiety Inventory (STAI) designed to assess both long-term and short-term anxiety features; and the Whitaker Index of Schizophrenic Thinking (WIST) designed to measure the type of thought impairment that differentiates between schizophrenic and "normal" thought processes.

While certain objective personality measures were developed for and have been used widely in clinical setting, several inventories have gained use in research and other normal range assessment settings. Many researchers utilize inven-

tories that measure the Big Five (or Five Factor Model, FFM) personality traits: Neuroticism (N), Extroversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). The NEO Personality Inventory (NEO-PI) assesses N, E, and O and has been used as an index of the Big Five in research. The Multidimensional Personality Questionnaire (MPQ) assesses the broad personality domains of Positive Affectivity, Negative Affectivity, and Constraint. The MPQ contains two validity indexes (VRIN and TRIN) common to the MMPI-2, which make the MPQ appealing to many researchers. Other objective personality instruments that have been used widely in research include the Personality Research Form (PRF), a measure of several psychosocial needs; the 16PF, which measures global factors similar to the Big Five; and the California Psychological Inventory (CPI), designed to assess "everyday" concepts about personality.

Finally, professionals working in the fields of educational/vocational assessment and personnel selection have found use for several objective personality measures. The 16PF, the CPI, and, most commonly, the MMPI/MMPI-2 have been used in both of these settings. Objective personality instruments give professionals information about an individual's personality which would not be obtained through standard applications or interviews. Thus, professionals can use objective personality inventories as an efficient method of obtaining more information to aid them in their task of advising clients about educational/vocational decisions or advising employers in the selection of personnel.

Whether one needs a tool to assess adult personality in clinical, research, or industry settings, there is probably an objective personality inventory to fit the bill. Many of today's objective inventories offer the efficiency of providing a comprehensive assessment of personality functioning, and some even provide computerized interpretation of the personality profile. Perhaps most importantly, most of the objective personality inventories available commercially are standardized instruments that can assess adult personality validly and reliably, which means these instruments can be used over the course of a client's treatment to document goals for change—a feature that is becoming increasingly important to clinicians in this era of managed healthcare.

4.14.6 REFERENCES

- Allard, C., & Mishara, B. L. (1995). Individual differences in stimulus intensity modulation and its relationship to two styles of depression in older adults. *Psychology and Aging, 10*, 395–403.

- Alterman, A. I., Zaballero, A. R., Lin, M. M., Siddiqui, N., Brown, L. S., Jr., Rutherford, M. J., & McDermott, P. A. (1995). Personality Assessment Inventory scores of lower-socioeconomic African American and Latino methadone maintenance patients. *Assessment*, 2, 91–100.
- Anderson, W. F. (1949). *Predicting success in nurses training*. Unpublished master's thesis, University of Nebraska, Lincoln, NE.
- Appleby, T. L., & Haner, C. F. (1956). MMPI profiles of a college faculty group. *Proceedings of the Iowa Academy of Sciences*, 53, 605–609.
- Applezweig, M. H. (1953). Educational levels and Minnesota Multiphasic profiles. *Journal of Clinical Psychology*, 9, 340–344.
- Arbisi, P., & Ben-Porath, Y. S. (1995). An MMPI-2 infrequency scale for use with psychopathological populations: The Infrequency–Psychopathology Scale, F (p). *Psychological Assessment*, 7, 424–431.
- Archer, R. P., Griffin, R., & Aiduk, R. (1995). Clinical correlates for ten common code types. *Journal of Personality Assessment*, 65, 391–408.
- Assendorp, J. B. (1992). Beyond stability: Predicting inter-individual differences in intra-individual change. *European Journal of Personality*, 6, 103–117.
- Bailey, D. M. (1988). Occupational therapy administrators and clinicians: Differences in demographics and values. *Occupational Therapy Journal of Research*, 8, 299–315.
- Bailey, S. E., & Ross, S. R. (1996, May). *The effects of simulated faking on the five-factor (NEO-FFI)*. Paper given at the 68th Annual Meeting of Midwestern Psychological Association, Chicago.
- Barger, B., & Hall, E. (1964). Personality patterns and achievement in college. *Educational and Psychological Measurement*, 24, 339–346.
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1–26.
- Barthol, R. P., & Kirk, B. A. (1956). The selection of graduate students in public health education. *Journal of Applied Psychology*, 40, 159–163.
- Bartram, D. (1992). The personality of UK managers: 16PF norms for short-listed applicants. *Journal of Occupational and Organizational Psychology*, 65, 159–172.
- Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8, 77–100.
- Ben-Porath, Y. S., Butcher, J. N., & Graham, J. R. (1991). Contribution of the MMPI-2 scales to the differential diagnosis of schizophrenia and major depression. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 3, 634–640.
- Ben-Porath, Y. S., McCully, E., & Almagor, M. (1993). Incremental validity of the MMPI-2 Content Scales in the assessment of personality and psychopathology by self-report. *Journal of Personality Assessment*, 61, 557–575.
- Beutler, L. E., Nussbaum, P. D., & Meredith, K. E. (1988). Changing patterns of police officers. *Professional Psychology: Research and Practice*, 19, 503–507.
- Beutler, L. E., Storm, A., Kirkish, P., Scogin, F., & Gaines, J. A. (1985). Parameters in the prediction of police officer performance. *Professional Psychology: Research and Practice*, 16, 324–335.
- Blake, D. D., Penk, W. E., Mori, D. L., Kleespies, P. M., Walsh, S. S., & Keane, T. M. (1992). Validity and clinical scale comparisons between MMPI and MMPI-2 with psychiatric inpatients. *Psychological Reports*, 70, 323–332.
- Boyle, G. J., & Lennon, T. J. (1994). Examination of the reliability and validity of the Personality Assessment Inventory. *Journal of Psychopathology and Behavioral Assessment*, 16, 173–187.
- Burbeck, E., & Furnham, A. (1985). Police officer selection: A critical review of the literature. *Journal of Police Science and Administration*, 13, 58–69.
- Burch, J. W. (1995). Typicality range deficit in schizophrenics' recognition of emotion in faces. *Journal of Clinical Psychology*, 51, 140–150.
- Burger, G. K., & Cross, D. T. (1979). Personality types as measured by the California Psychological Inventory. *Journal of Consulting and Clinical Psychology*, 47, 65–71.
- Burgess, E. (1956). Personality factors of over- and under-achievers in engineering. *Journal of Educational Psychology*, 47, 89–99.
- Bush, A. J., & Lucas, G. H. (1988). Personality profiles of marketing vs. R&D managers. *Psychology and Marketing*, 5, 17–32.
- Butcher, J. N. (1989). *MMPI-2 scale correlates*. MMPI-2 Workshop Materials. Minneapolis MN: University of Minnesota Press.
- Butcher, J. N. (1991). Screening for psychopathology: Industrial applications of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2). In J. Jones, B. D. Steffey, & D. Bray (Eds.), *Applying psychology in business: The manager's handbook*. Boston: Lexington.
- Butcher, J. N. (1994). Psychological assessment of airline pilot applicants with the MMPI-2. *Journal of Personality Assessment*, 62, 31–44.
- Butcher, J. N. (1995). Clinical use of computer-based personality test reports. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 78–94). New York: Oxford University Press.
- Butcher, J. N., Berah, E., Ellertsen, B., Miach, P., Lim, J., Nezami, E., Pancheri, P., Derksen, J., & Almagor, M. (1998). Objective personality assessment: Computer-based MMPI-2 interpretation in international clinical settings. In C. Belar (Ed.), *Comprehensive clinical psychology: Sociocultural and individual differences*. New York: Elsevier.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
- Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 Content Scales*. Minneapolis, MN: University of Minnesota Press.
- Butcher, J. N., & Han, K. (1995). Development of an MMPI-2 scale to assess the presentation of self in a superlative manner: The S Scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment*, (Vol. 10, pp. 25–50). Hillsdale, NJ: Erlbaum.
- Butcher, J. N., Jeffrey, T., Cayton, T. G., Colligan, S., DeVore, J., & Minnegawa, R. (1990). A study of active duty military personnel with the MMPI-2. *Military Psychology*, 2, 47–61.
- Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology*, 47, 87–111.
- Butcher, J. N., & Williams, C. L. (1992). *MMPI-2 and MMPI-A: Essentials of clinical interpretation*. Minneapolis, MN: University of Minnesota Press.
- Cattell, R. B. (1963). Personality, role, mood, and situation perception: A unifying theory of modulators. *Psychological Review*, 70, 1–18.
- Cattell, R. B., Cattell, A. K. S., & Cattell, H. E. P. (1993). *Sixteen Personality Factor Questionnaire* (5th ed.). Champaign, IL: Institute for Personality and Ability Testing.
- Centi, P. (1962). Personality factors related to college success. *Journal of Educational Research*, 55, 187–188.
- Cerf, A. Z. (1947). Personality inventories. In J. P. Guilford

- (Ed.), *Printed classification tests*. Washington, DC: Army Air Force Aviation Psychology Program Research Reports.
- Cerling, G. L. (1983). Selection of lay counselors for a church counseling center. *Journal of Psychology and Christianity*, 2, 67-72.
- Chakrabarti, P. K., & Kundu, R. (1984). Personality profiles of management personnel. *Psychological Studies*, 29, 143-146.
- Chick, D., Martin, S. K., Nevels, R., & Cotton, C. R. (1994). Relationship between personality disorders and clinical symptoms in psychiatric inpatients as measured by the Millon Clinical Multiaxial Inventory. *Psychological Reports*, 74, 331-336.
- Chick, D., Sheaffer, C. I., Goggin, W. C., & Sison, G. F. (1993). The relationship between MCMI personality scales and clinician generated *DSM-III* personality disorder diagnoses. *Journal of Personality Assessment*, 61, 264-276.
- Clark, D. L. (1964). Exploring behavior in men's residence halls using the MMPI. *Personnel Guidance Journal*, 43, 249-251.
- Clark, J. H. (1953). Grade achievement of female college students in relation to non-intellective factors: MMPI items. *Journal of Social Psychology*, 37, 275-281.
- Clark, M. E. (1996). MMPI-2 negative treatment indicators content and content component scales: Clinical correlates and outcome prediction for men with chronic pain. *Psychological Assessment*, 8, 32-47.
- Conley, J. J. (1985). Longitudinal stability of personality traits: A multi trait-multi method-multi occasion analysis. *Journal of Personality and Social Psychology*, 49, 1266-1282.
- Cooper, C. L., & Green, M. D. (1976). Coping with occupational stress among Royal Air Force personnel on isolated island bases. *Psychological Reports*, 39, 731-734.
- Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4, 5-13.
- Crookes, T. G., & Buckley, S. J. (1976). Lie score and insight. *Irish Journal of Psychology*, 3, 134-136.
- Deniston, W. M., & Ramanaiah, N. V. (1993). California Psychological Inventory and the five-factor model of personality. *Psychological Reports*, 73, 491-496.
- Dinning, W. D., & Evans, R. G. (1977). Discriminant and convergent validity of the SCL-90 in psychiatric inpatients. *Journal of Personality Assessment*, 41, 304-310.
- Dunn, W. S., Mount, M. K., Barrick, M. R., & Ones, D. S. (1995). Relative importance of personality and general mental ability in managers' judgments of applicant qualifications. *Journal of Applied Psychology*, 80, 500-509.
- Dyer, J. B., Sajwaj, T. E. G., & Ford, T. W. X. (1993, March). *MMPI-2 normative and comparative data for nuclear power plant personnel who were approved or denied security clearances for psychological reasons*. Paper presented at the 28th Annual Symposium on Recent Developments in the Use of the MMPI/MMPI-2, St. Petersburg, FL.
- Egeland, B., Erickson, M., Butcher, J. N., & Ben-Porath, Y. S. (1991). MMPI-2 profiles of women at risk for child abuse. *Journal of Personality Assessment*, 57, 254-263.
- Evans, R. G., & Dinning, W. D. (1980). A validation of Forms A and B of the Whitaker Index of Schizophrenic Thinking. *Journal of Personality Assessment*, 44, 416-419.
- Fabricatore, J., Azen, S. P., Schoentgen, S., & Snibbe, H. (1978). Predicting performance of police officers using the Sixteen Personality Factor Questionnaire. *American Journal of Community Psychology*, 6, 63-70.
- Fekken, G. C. (1985). The Whitaker Index of Schizophrenic Thinking. In D. J. Keyser & R. C. Sweetland (Eds.), *Test critiques* (Vol. III, pp. 717-725). Kansas City, KS: Test Corporation Of America.
- Fekken, G. C., & Holden, R. R. (1991). The construct validity of person reliability. *Personality and Individual Differences*, 12, 69-77.
- Ferris, G. R., Bergin, T. G., & Wayne, S. J. (1988). Personal characteristics, job performance, and absenteeism of public school teachers. *Journal of Applied Social Psychology*, 18, 552-563.
- Finn, S. E. (1986). Stability of personality self-ratings over 30 years: Evidence for an age/cohort interaction. *Journal of Personality and Social Psychology*, 50, 813-818.
- Fishkin, S. M., Lovallo, W. R., & Pishkin, V. (1977). Relationship between schizophrenic thinking and MMPI for process and reactive patients. *Journal of Clinical Psychology*, 33, 116-119.
- Flynn, P. M. (1995). Issues in the assessment of personality disorder and substance abuse using the Millon Clinical Multiaxial Inventory (MCMI-II). *Journal of Clinical Psychology*, 51, 415-421.
- Frick, J. W. (1955). Improving the prediction of academic achievement by use of the MMPI. *Journal of Applied Psychology*, 39, 49-52.
- Fulkerson, S. C., Freud, S. L., & Raynor, G. H. (1958, February). The use of the MMPI in psychological evaluation of pilots. *Aviation Medicine* 122-128.
- Fulkerson, S. C., & Sells, S. B. (1958). Adaptation of the MMPI for aeromedical practice norms for military pilots. *USAF School of Aviation Medicine Reports*, 58-128.
- Furnham, A. (1991). Personality and occupational success: 16PF correlates of cabin crew performance. *Personality and Individual Differences*, 12, 87-90.
- Garetz, F. K., & Tierney, R. W. (1962). Personality variables in army officer candidates. *Military Medicine*, 127, 669-672.
- Geist, C. R., & Boyd, S. T. (1980). Personality characteristics of Army helicopter pilots. *Perceptual Motor Skills*, 51(1), 253-254.
- Gorner, A. B. (1970). MMPI and MMPI scores, correlations and analysis for military aircrew population. *British Journal of Social and Clinical Psychology*, 9, 164-170.
- Gottfredson, G. D., Jones, E. M., & Holland, J. L. (1993). Personality and vocational interests: The relation of Holland's six interest dimensions to five robust dimensions of personality. *Journal of Counseling Psychology*, 40, 518-524.
- Gough, H. G. (1957). *Manual for the California Psychological Inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Graham, J. R., & Ben-Porath, Y. S. (1990, June). *Congruence between the MMPI and MMPI-2*. Paper given at the 25th Annual Symposium on Recent Developments in the Use of the MMPI/MMPI-2, Minneapolis, MN.
- Graham, J. R., & Butcher, J. N. (1988, March). *Differentiating schizophrenic and major affective disorders with the revised form of the MMPI*. Paper presented at the 23rd Annual Symposium on Recent Developments in the Use of the MMPI, St. Petersburg, FL.
- Grigoriadis, S. (1993). *Sources of inconsistency on tests of psychopathology*. Unpublished doctoral dissertation, Queen's University, Kingston, ON.
- Hargrave, G. E., & Hiatt, D. (1987, May). *Use of the MMPI to predict aggression in law enforcement officer applicants*. Paper presented at the 22nd Annual Symposium on Recent Developments in the Use of the MMPI, Seattle, WA.
- Hargrave, G. E., & Hiatt, D. (1989). Use of the California Psychological Inventory in law enforcement officer selection. *Journal of Personality Assessment*, 53, 267-277.
- Hartman, B. J. (1987). Psychological screening of law

- enforcement candidates. *American Journal of Forensic Psychology*, 5, 5-10.
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology*, 10, 249-254.
- Helson, R., & Moane, G. (1987). Personality change in women from college to midlife. *Journal of Personality and Social Psychology*, 53, 176-186.
- Henney, A. S. (1975). Personality characteristics of a group of industrial managers. *Journal of Occupational Psychology*, 48, 65-67.
- Hills, H. A. (1995). Diagnosing personality disorders: An examination of the MMPI-2 and MCMI-II. *Journal of Personality Assessment*, 65, 21-34.
- Hindelang, M. J. (1972). The relationships of self-reported delinquency to scales of the CPI and MMPI. *Journal of Criminal Law, Criminology, and Police Science*, 63, 75-81.
- Hjemboe, S., & Butcher, J. N. (1991). Couples in marital distress: A study of demographic and personality factors as measured by the MMPI-2. *Journal of Personality Assessment*, 57, 216-237.
- Hoffman, R. G., & Davis, G. L. (1995). Prospective validity study: CPI work orientation and managerial potential scales. *Educational and Psychological Measurement*, 55, 881-890.
- Holden, R. R., Fekken, G. C., Reddon, J. R., Helmes, E., & Jackson, D. N. (1988). Clinical reliabilities and validities of the Basic Personality Inventory. *Journal of Consulting and Clinical Psychology*, 56, 766-768.
- Husband, S. D., & Iguchi, M. (1995). Comparison of MMPI-2 and MMPI clinical scales and high point scores among methadone maintenance clients. *Journal of Personality Assessment*, 64, 371-375.
- Inch, R., & Crossley, M. (1993). Diagnostic utility of the MCMI-I and MCMI-II with psychiatric outpatients. *Journal of Clinical Psychology*, 49, 358-366.
- Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Port Huron, MI: Research Psychologists Press.
- Jackson, D. N. (1989). *Basic Personality Inventory manual*. Port Huron, MI: Research Psychologists Press.
- Jackson, D. N., & Messick, S. (1971). *The Differential Personality Inventory*. London, ON: Authors.
- Jaffe, L. T., & Archer, R. P. (1987). The prediction of drug use among college students from MMPI, MCMI, and sensation seeking scales. *Journal of Personality Assessment*, 51, 243-253.
- Jennings, L. S. (1948). Minnesota Multiphasic Personality Inventory; differentiation of psychologically good and poor combat risks among flying personnel. *Journal of Aviation Medicine*, 19, 222.
- Johnson, R. W., Flammer, D. P., & Nelson, J. G. (1975). Multiple correlations between personality factors and SVIB occupational scales. *Journal of Counseling Psychology*, 22, 217-223.
- Keilen, W. G., & Bloom, L. J. (1986). Child custody evaluation practices: A survey of experienced professionals. *Professional Psychology: Research and Practice*, 17, 338-346.
- Keller, L. S., & Butcher, J. N. (1991). *Use of the MMPI-2 with chronic pain patients*. Minneapolis, MN: University of Minnesota Press.
- Khan, F. I., Welch, T., & Zillmer, E. (1998). MMPI-2 profiles of battered women in transition. *Journal of Personality Assessment*, 60, 100-111.
- Kline, P., & Lapham, S. L. (1992). Personality and faculty in British universities. *Personality and Individual Differences*, 13, 855-857.
- Knight, R. A., Epstein, B., & Zielony, R. D. (1980). The validity of the Whitaker Index of Schizophrenic Thinking. *Journal of Clinical Psychology*, 36, 632-639.
- Koss, M. P., & Butcher, J. N. (1973). A comparison of psychiatric patients' self-report with other sources of clinical information. *Journal of Research in Personality*, 7, 225-236.
- Krueger, R. F., Schmutte, P. S., Caspi, A., Moffitt, T. E., Campbell, K., & Silva, P. A. (1994). Personality traits are linked to crime among men and women: Evidence from a birth cohort. *Journal of Abnormal Psychology*, 103, 328-338.
- Kuhne, A., Orr, S., & Baraga, E. (1993). Psychometric evaluation of post-traumatic stress disorder: The Multi-dimensional Personality Questionnaire as an adjunct to the MMPI. *Journal of Clinical Psychology*, 49, 218-225.
- Lardent, C. L. (1991). Pilots who crash: Personality constructs underlying accident prone behavior of fighter pilots. *Multivariate Experimental Clinical Research*, 10, 1-25.
- Laufer, W. S., Skoog, D. K., & Day, J. M. (1982). Personality and criminality: A review of the California Psychological Inventory. *Journal of Clinical Psychology*, 38, 562-573.
- Lavin, P. F., Chardos, S. P., Ford, W. T., & McGee, R. K. (1987). The MMPI profiles of troubled employees in relation to nuclear power plant personnel norms. *Transactions of the American Nuclear Society*, 54, 146-147.
- Lawrence, R. A. (1984). Police stress and personality factors: A conceptual model. *Journal of Criminal Justice*, 12, 247-263.
- Leslie, B. A., Landmark, J., & Whitaker, L. C. (1984). The Whitaker Index of Schizophrenic Thought (WIST) and thirteen systems for diagnosing schizophrenia. *Journal of Clinical Psychology*, 40, 636-648.
- Lorr, M., & Strack, S. (1994). Personality profiles of police candidates. *Journal of Clinical Psychology*, 50, 200-207.
- Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935-1982. *American Psychologist*, 39, 451-454.
- Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, 1, 477-482.
- MacCabe, S. P. (1987). Millon Clinical Multiaxial Inventory. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques Compendium* (pp. 304-315). Kansas City, KS: Test Corporation Of America.
- MacLennan, R. N. (1991). *Personality Research Form annotated research bibliography*. Regina, AL: University of Regina, Department of Psychology.
- McCrae, R. R., & Costa, P. T., Jr. (1990). *Personality in adulthood*. New York: Guilford.
- Megargee, E. I. (1972). *The California Psychological Inventory handbook*. San Francisco: Jossey-Bass.
- Millon, T. (1977). *Manual for the Millon Clinical Multiaxial Inventory*. Minneapolis, MN: National Computer Systems.
- Millon, T. (1987). *Manual for the Millon Clinical Multiaxial Inventory-II*. Minneapolis, MN: National Computer Systems.
- Millon, T. (1994). *Manual for the Millon Clinical Multiaxial Inventory-III*. Minneapolis, MN: National Computer Systems.
- Millon, T., & Davis, R. (1995). Putting Humpty Dumpty together again: Using the MCMI in psychological assessment. In L. E. Beutler & M. R. Berren (Eds.), *Integrative assessment of adult personality* (pp. 240-279). New York: Guilford.
- Morey, L. (1991). *Personality Assessment Inventory: Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Murray, H. A. (1938). *Explorations in personality*. Cambridge, MA: Harvard University Press.
- Patrick, J. (1993). Validation of the MCMI-I Borderline Personality Disorder scale with a well-defined criterion sample. *Journal of Clinical Psychology*, 49, 28-32.
- Paunonen, S. V., Jackson, D. N., Trzebinski, J., & Fosterling, F. (1992). Personality structures across

- cultures: A multi method evaluation. *Journal of Personality and Social Psychology*, 62, 447–456.
- Payne, F. D., & Wiggins, J. S. (1972). MMPI profile types and the self-report of psychiatric patients. *Journal of Abnormal Psychology*, 79, 1–8.
- Payne, R. W. (1978). Review of Whitaker Index of Schizophrenic Thought. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1146–1147). Highland Park, NJ: Gryphon Press.
- Piedmont, R. L. (1993). A longitudinal analysis of burnout in the health care setting: The role of personal dispositions. *Journal of Personality Assessment*, 61, 457–473.
- Pishkin, V., Lovallo, W. R., & Bourne, L. E. (1986). Thought disorder and schizophrenia: Isolating and timing a mental event. *Journal of Clinical Psychology*, 42, 417–424.
- Reynolds, C. R. (1992). Review of the Millon Clinical Multiaxial Inventory-II. In J. J. Kramer & J. C. Conoley (Eds.), *The eleventh mental measurements yearbook* (pp. 533–535). Lincoln, NE: Buros Institute of Mental Measurements.
- Saxe, S. J., & Reiser, M. (1976). A comparison of three police applicant groups using the MMPI. *Journal of Police Science and Administration*, 4, 419–425.
- Schinka, J. A. (1995). Personality Assessment Inventory scale characteristics and factor structure in the assessment of alcohol dependency. *Journal of Personality Assessment*, 64, 101–111.
- Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, 78, 966–974.
- Schretlen, D. (1988). The use of psychological tests to identify malingering symptoms of mental disorder. *Clinical Psychology Review*, 8, 451–476.
- Schuerger, J. M., Zarrella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology*, 56, 777–783.
- Scogin, F., & Beutler, L. (1986). Psychological screening of law enforcement candidates. In P. A. Keller & L. G. Ritt (Eds.), *Innovations in clinical practice* (Vol. 5, pp. 317–330). Sarasota, FL: Professional Resources Exchange.
- Scogin, F., & Reiser, M. (1976). A comparison of three police applicant groups using the MMPI. *Journal of Police Science and Administration*, 4, 419–425.
- Soldz, S., Budman, S., Demby, A., & Merry, J. (1993). Diagnostic agreement between the Personality Disorder Examination and the MCMI-II. *Journal of Personality Assessment*, 60, 486–499.
- Spielberger, C. D. (1977). Anxiety, theory and research. In B. Wolman (Ed.) *International encyclopedia of neurology, psychiatry, psychoanalysis, and psychology*. New York: Human Sciences Press.
- Spielberger, C. D. (1976). Stress and anxiety and cardiovascular disease. *Journal of South Carolina Medical Association* (Supplement 15), 72, 15–22.
- Spielberger, C. D. (1983). *Manual for the State Trait Anxiety Inventory: STAI (form Y)*. Palo Alto, CA: Consulting Psychologists Press.
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. D. (1970). *The STAI: Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Spielberger, C. D., Ritterband, L. M., Sydeman, S. J., Reheiser, E. C., and Unger, K. K. (1995). Assessment of emotional states and personality traits: Measuring psychological vital signs. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 43–58). New York: Oxford University Press.
- Tellegen, A. *Brief manual for the Differential Personality Questionnaire*. Unpublished manuscript.
- Topp, B. W., & Kardash, C. A. (1986). Personality, achievement, and attrition: Validation in a multiple-jurisdiction police academy. *Journal of Police Science and Administration*, 14, 234–241.
- Walsh, W. B. (1974). Consistent occupational preferences and personality. *Journal of Vocational Behavior*, 4, 145–153.
- Watkins, C. E. (1996). On Hunsley, Harangue, and Hoopla. *Professional Psychology: Research and Practice*, 27, 316–318.
- Weed, N. C., Butcher, J. N., Ben-Porath, Y. S., & McKenna, T. (1992). New measures for assessing alcohol and drug abuse with the MMPI-2: The APS and AAS. *Journal of Personality Assessment*, 58, 389–404.
- Whitaker, L. C. (1973). *The Whitaker Index of Schizophrenic Thinking*. Los Angeles: Western Psychological Services.
- Whitaker, L. C. (1980). *Objective measurement of schizophrenic thinking: A practical and theoretical guide to the Whitaker Index of Schizophrenic Thinking*. Los Angeles: Western Psychological Services.
- Whitbourne, S. K., Zuschlag, L. B., Elliot, L. B., & Waterman, A. S. (1992). Psychosocial development in adulthood: A 22-year sequential study. *Journal of Personality and Social Psychology*, 63, 260–271.
- Wiggins, J. S. (1969). Content dimensions in the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications* (pp. 127–180).
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.
- Williams, J. E., Munick, M. L., Saiz, J. L., & Formy-Duval, D. L. (1995). Psychological importance of the “Big Five”: Impression formation and context effects. *Personality and Social Psychology Bulletin*, 21, 818–826.
- Windle, C. (1954). Test-retest effect on personality questionnaires. *Educational and Psychological Measurement*, 14, 617–633.

4.15

Projective Assessment of Children and Adolescents

IRVING B. WEINER and KATHRYN KUEHNLE

University of South Florida, Tampa, FL, USA

4.15.1 INTRODUCTION	432
4.15.2 OBJECTIVITY AND SUBJECTIVITY IN PERSONALITY ASSESSMENT METHODS	432
4.15.3 STRUCTURE AND AMBIGUITY IN PROJECTIVE TECHNIQUES	433
4.15.4 VALUE OF PROJECTIVE ASSESSMENT	435
4.15.4.1 <i>Conceptual Basis</i>	435
4.15.4.2 <i>Empirical Basis</i>	436
4.15.5 UTILITY OF PROJECTIVE ASSESSMENT	436
4.15.6 APPLICABILITY OF PROJECTIVE METHODS TO CHILDREN AND ADOLESCENTS	437
4.15.7 REVIEW OF PROJECTIVE ASSESSMENT METHODS	439
4.15.7.1 <i>Rorschach Inkblot Method</i>	439
4.15.7.1.1 <i>Administration and scoring</i>	439
4.15.7.1.2 <i>Psychometric foundations</i>	439
4.15.7.1.3 <i>Clinical utility</i>	440
4.15.7.2 <i>Thematic Apperception Test</i>	440
4.15.7.3 <i>Administration and scoring</i>	441
4.15.7.3.1 <i>Psychometric foundations</i>	441
4.15.7.3.2 <i>Clinical utility</i>	442
4.15.7.4 <i>Children's Apperception Test</i>	442
4.15.7.4.1 <i>Administration and scoring</i>	443
4.15.7.4.2 <i>Psychometric foundations</i>	443
4.15.7.4.3 <i>Clinical utility</i>	443
4.15.7.5 <i>Roberts Apperception Test for Children</i>	444
4.15.7.5.1 <i>Administration and scoring</i>	444
4.15.7.5.2 <i>Psychometric foundations</i>	444
4.15.7.5.3 <i>Clinical Utility</i>	445
4.15.7.6 <i>Tell-me-a-story</i>	445
4.15.7.6.1 <i>Administration and scoring</i>	446
4.15.7.6.2 <i>Psychometric foundations</i>	446
4.15.7.6.3 <i>Clinical utility</i>	447
4.15.7.7 <i>Draw-a-person</i>	447
4.15.7.7.1 <i>Administration and scoring</i>	448
4.15.7.7.2 <i>Psychometric foundations</i>	448
4.15.7.7.3 <i>Clinical utility</i>	449
4.15.7.8 <i>House-tree-person</i>	449
4.15.7.8.1 <i>Administration and scoring</i>	450
4.15.7.8.2 <i>Psychometric foundations</i>	450
4.15.7.8.3 <i>Clinical utility</i>	450
4.15.7.9 <i>Kinetic Family Drawing</i>	450
4.15.7.9.1 <i>Administration and scoring</i>	451

4.15.7.9.2 Psychometric foundations	451
4.15.7.9.3 Clinical utility	452
4.15.7.10 Sentence Completion Methods	452
4.15.7.10.1 Administration and scoring	452
4.15.7.10.2 Psychometric foundations	453
4.15.7.10.3 Clinical utility	453
4.15.8 FUTURE DIRECTIONS	454
4.15.9 SUMMARY	454
4.15.10 REFERENCES	455

4.15.1 INTRODUCTION

Projection, as first formulated by Freud (1962) and later elaborated in his presentation of the Schreber case (Freud, 1958), consists of attributing one's own characteristics to external objects or events without adequate justification or conscious awareness of doing so. Frank (1939) suggested that personality tests in which there is relatively little structure induce a subject to "project upon that plastic field . . . his private world of personal meanings and feelings" (pp. 395, 402). By linking the concept of projection to the response process in such measures as the Rorschach Inkblot Method and the Thematic Apperception Test (TAT), Frank gave birth to the so-called projective hypothesis in personality assessment. His observations about what he called "projection measures" led to the Rorschach, the TAT, and other assessment methods involving some ambiguity being routinely designated as projective tests.

This chapter begins with some observations concerning the nature of objectivity and subjectivity in personality assessment methods and the role of structure and ambiguity in different types of projective techniques. It then turns to the value and utility of projective assessment and the applicability of projective methods in clinical work with young people. Information is given on the composition, administration, scoring, psychometric foundations, and clinical utility of nine projective techniques widely used in assessing children and adolescents.

4.15.2 OBJECTIVITY AND SUBJECTIVITY IN PERSONALITY ASSESSMENT METHODS

As a legacy of Frank's projective hypothesis, tests designated as projective methods came to be regarded as subjective in nature and hence quite different from objective methods, such as self-report inventories. This presumed subjectivity of projective methods fostered a commonplace conviction that these methods are inherently less scientific and less valid than

objective methods. In actuality, however, ambiguity is a dimensional rather than a categorical characteristic of tests, and there is little basis for regarding projective methods as inherently unscientific and invalid or as sharply distinct from objective measures.

Being scientific does not inhere in the nature of a method or instrument, whether subjective or not, but only in whether it can be studied scientifically. When projective tests are used to generate personality descriptions that can be independently and reliably assessed for their accuracy, they function as a scientific procedure.

Likewise, the validity of a test inheres not in its nature, but rather in the extent to which it generates significant correlations with personality characteristics or behaviors it can identify or predict. Abundant research attests that projective methods, when properly used, can yield valid inferences (Hibbard et al., 1994; Hibbard, Hilsenroth, Hibbard, & Nash, 1995; Parker, Hanson, & Hunsley, 1988; Weiner, 1996).

Regarding sharp distinctions between projective and objective measures, the subject's task on commonly used personality tests injects considerable objectivity into many projective methods and substantial subjectivity into most objective methods. For example, responses on the Rorschach, which is the most widely used projective measure, are routinely inquired by asking subjects "Where did you see it?," which is a concrete, unambiguous request for a specific item of objective information. When subjects reply that they used the whole blot for a percept, their response is given a location choice code of W, which is an objective and unambiguous procedure on which coders achieve virtually 100% agreement. Numerous other features of how subjects choose to look at the inkblots can also be objectively coded with good inter-rater reliability, such as the percentage of responses given to the multicolored cards (affective ratio), the number of commonly given percepts reported (populars), and the kinds of objects the blots are said to resemble (people, animals, etc.) (Exner, 1991, pp. 459–460; McDowell & Acklin, 1996; Seaton & Allen, 1996).

Elements of objectivity mark the interpretation as well as the coding of many Rorschach variables. In the case of whole responses, for example, an unusual preponderance of W in a record correlates with objectively observable tendencies to attend to experience in a global fashion; a low affective ratio identifies an inclination to withdraw from emotionally charged situations; a small number of popular responses correlates with behavioral manifestations of unconventionality; and numerous human percepts is associated with an active interest in people. In these and many other ways, Rorschach responses identify personality characteristics through an objective process of coding response features and relating these coded features to their known corollaries in observable behavior.

There are aspects of Rorschach interpretation that may be highly subjective, especially when inferences are drawn from the thematic imagery subjects produce when they associate to the inkblots along with describing them. Moreover, most other projective methods have not been as extensively codified as the Rorschach and depend more on qualitative than quantitative analysis. The present point is merely that the basic nature of projective methods does not preclude their being codified and interpreted to some extent along objective lines.

Turning now to aspects of subjectivity in objective methods, consider the uncertainty that characterizes many items in the most widely used objective measure of personality, the adult and adolescent forms of the Minnesota multiphasic personality inventory (MMPI-2/MMPI-A). Although MMPI-2/MMPI-A instructions to respond true or false are unambiguous and the coding of these responses is completely objective, Weiner (1993) has previously called attention to the idiography that is embedded in asking subjects to interpret such items as "I often lose my temper." Items of this type provide no benchmarks for the frequency of "often" or for what constitutes loss of temper. In many instances, consequently, responses to self-report items involve subjectivity on the part of respondents, who must define for themselves what certain terms mean before they can decide how to answer.

Subjectivity influences the interpretation as well as the response process in objective assessment. Granted, the hallmark of objective tests is an extensive array of quantified scale scores having empirically demonstrated behavioral correlates. Nevertheless, the interpretation of self-report measures in clinical practice typically goes beyond identifying known corollaries of scale scores to include consideration of complex patterns of interaction among these

scores and between the test profile and aspects of a subject's clinical history, interview behavior, and performance on other tests. Some of these complex interactions have been examined empirically, such as various two- and three-point codes on the MMPI-2/MMPI-A, but many have not.

This is not to say that the MMPI-2/MMPI-A and other self-report instruments are basically subjective in nature or that they derive their utility primarily from clinical judgment. The point is merely that, just as projective instruments are not entirely subjective, self-report methods are not completely objective, but instead involve some aspects of ambiguity in how subjects respond to them and how examiners interpret them.

It is for this reason that ambiguity is not a categorical function that characterizes some tests called projective measures, but not others called objective measures. Instead, ambiguity is a dimensional function that characterizes most tests to some degree, in relation to how structured they are. Generally speaking, objective tests are more structured than projective tests and therefore less ambiguous; projective tests are generally less structured than objective tests and hence more ambiguous; and there is no sharp objective/subjective dichotomy between relatively structured and relatively unstructured instruments.

4.15.3 STRUCTURE AND AMBIGUITY IN PROJECTIVE TECHNIQUES

Projective techniques comprise inkblot methods, story-telling methods, figure drawing methods, and sentence completion methods. In addition to being less structured than objective measures, these four types of projective methods differ from each other in their degree of ambiguity and in whether their ambiguity resides mainly in their stimuli or in their instructions.

Thus in the case of the Rorschach inkblot method, subjects are asked to look at relatively ambiguous stimuli but are given fairly specific instructions to indicate what they see, where they see it, and what makes it look as it does. Story-telling methods such as the TAT involve showing subjects real pictures that are much less ambiguous than inkblots; however, by using general instructions ("Tell me a story") and open-ended questions ("What will happen next?"), examiners provide only minimal guidance in how subjects should respond. If the Rorschach instructions were "Tell me a story about this inkblot," the Rorschach would be more ambiguous and more of a projective test

than it is. If the TAT instructions were "Tell me what you see here," the TAT would lose most of its ambiguity and function only barely as a projective test.

Figure drawing techniques use no stimuli at all, save a blank piece of paper, and provide little guidance to subjects, other than some instructions concerning the figures to be drawn (e.g., yourself, a family). Sentence completion methods, in common with story-telling techniques, call for subjects to provide thematic content in response to real and relatively unambiguous test stimuli. Unlike story-telling techniques, however, sentence completion methods do not ordinarily involve querying subjects about their responses or encouraging them to elaborate those that are brief or unrevealing. Thus a stem of "I AM" may be completed with "a happy person," in which case some subjectivity has been allowed to enter the response, or simply with "here," in which case only a completely objective response has been given. On balance, figure drawing techniques are the most ambiguous of projective tests and sentence completion methods the least, with inkblot and story-telling techniques in between.

These differences in ambiguity among projective methods were originally noted by Stone and Dellis (1960), who proposed "a levels hypothesis" to take practical account of this variability. According to the levels hypothesis, the degree to which a test is structured is directly related to the level of conscious awareness at which it taps personality processes. The more structured and less ambiguous a test is, the more likely it is to yield information about relatively conscious and superficial levels of personality; conversely, the less structured and more ambiguous a test is, the more likely it is to provide information about deeper levels of personality and characteristics of which subjects themselves may not be consciously aware.

Research reported by Stone and Dellis (1960) and subsequently replicated by Murstein and Wolf (1970) provided empirical support for a relationship between the ambiguity of a test and its likelihood of measuring deeper levels of personality, especially in normally functioning persons. These findings mirrored the basic conception of TAT assessment articulated by Murray (1951), who regarded the virtue of the instrument as residing not in its revelations about what subjects are able and willing to say about themselves, but in what it conveys about personality characteristics: "the patient is unwilling to tell or is unable to tell because he is unconscious of them" (p. 577).

In addition to differing from objective tests and from each other in their degree of structure and ambiguity, individual projective measures

typically include both relatively objective and relatively subjective elements. As elaborated by Weiner (1977), the objective elements of projective test data involve structural features of the manner in which responses are formulated, whereas the subjective elements consist of thematic features of the imagery with which responses are embellished.

When projective test data are being interpreted objectively, structural aspects of the subject's responses, such as focusing on wholes and seeing numerous human figures on the Rorschach, are taken as being directly representative of similar behavioral tendencies in the person's life, that is, attending to experience globally and paying close attention to people. When projective test data are being interpreted subjectively, thematic imagery is taken as being indirectly symbolic of a subject's underlying needs, attitudes, conflicts, and concerns. Thus the Rorschach response of "Two girls who are really mad at each other fighting over something they both want" may identify a subject's experiencing peer or sibling rivalry, viewing social interactions as aggressive confrontations in which people are only concerned with what they can get for themselves, or feeling angry or resentful about being in such situations.

On story-telling measures, an example of a structural response feature is giving long stories, which can be objectively scored (by counting the number of words) and which provides a representative indication of inclinations to be verbose. As for subjectively interpreted features, a TAT story in which two people are described as about to separate, leaving one of them sad and lonely for the rest of his or her life, exemplifies thematic imagery that appears to symbolize concerns about suffering the loss of love objects and facing an unhappy future.

On figure drawing measures, which as previously noted are the most ambiguous of projective tests, structural features of the data are limited. Some variables, such as the size of figures drawn, how complete they are, and whether they are clothed, are objective facts that can usually be coded with good agreement. However, interpretation of such objective characteristics of figure drawings, as well as of subjective impressions of drawing qualities, is based mostly on their being symbolic rather than representative of behavior. Interpreting the way figures are drawn or placed is thus primarily thematic. For example, unusual emphasis on a particular part of the body may be interpreted as suggesting concern about functions associated with that part of the body, and a family drawing in which the self is located on one side of the page and the other family members are closely

grouped on the other side of the page may be interpreted as symbolizing feelings of isolation or rejection in the family setting.

In sentence completion responses, frequent self-referencing is an example of an objectively scorable, behaviorally representative structural index of tendencies to focus attention on oneself rather than others. Consider the difference between the completions "WHAT PAINS ME is seeing how many unfortunate people there are in the world" and "WHAT PAINS ME is not being able to get the things that I want." An accumulation of the latter as opposed to the former type of response is objectively representative of self-centeredness. At the same time, the thematic content of both completions suggests in a more subjective way certain underlying concerns, such as worries about the welfare of the human race in the first instances and feelings of being personally deprived in the second.

To bring these introductory observations full circle, the opportunities that projective methods create for subjects to project aspects of themselves into their responses has frequently led to their being associated with psychoanalytic theories of personality, in the context of which the notion of projection was first elaborated. However, there is no necessary relationship between psychoanalytic theory and projective testing, nor is there any reason for clinicians who conceptualize behavior in other ways to view projective methods as incompatible with their frame of reference. The basic principle underlying projective techniques is that something can be learned about people from sampling how they respond in ambiguous situations. This principle is not prisoner to any personality theory, and its utility transcends the theoretical persuasions of individual examiners. Inferences from projective data can be couched equally well in psychodynamic, behavioral, cognitive, and humanistic terms, and the use to which these inferences can be put depends less on theoretical differences in terminology than on the nature of the assessment issues being addressed.

4.15.4 VALUE OF PROJECTIVE ASSESSMENT

Projective test data provide valuable information about how people are likely to think, feel, and act that is difficult to obtain from objective assessment procedures. This contribution of projective methods to the personality assessment process has both a conceptual and an empirical basis.

4.15.4.1 Conceptual Basis

Because of their relatively unstructured nature, projective tests measure personality characteristics in subtle and indirect ways. Even those features of projective test data that can be objectively scored and interpreted involve responses that seldom have obvious meaning. Subjects in the process of responding usually have little awareness of the interpretive significance that attaches to their seeing numerous human figures on the Rorschach, giving long stories on the TAT, drawing themselves on the far side of the page from the rest of their family, or repetitively referring to "I" in their sentence completions; indeed, they may not even be aware of having responded in these ways.

By contrast, relatively structured objective tests measure personality characteristics in direct ways that often have obvious interpretive significance. Adolescents who answer "true" to such MMPI-A statements as "At times I feel like smashing things" and "I am easily downed in an argument" will usually have a good idea of what they are indicating about themselves.

The distinction between subtle, indirect measurement of personality characteristics with projective techniques and relatively direct assessment through questionnaire methods has been formulated by McClelland, Koestner, and Weinberger (1989) in terms of differences between self-attributed and implicit motives. According to McClelland et al., self-attributed motives are measured by self-report instruments and are influenced by social incentives in a person's external environment. Implicit motives, however, are measured by such indirect techniques as story-telling procedures and are influenced by the internal pleasure derived from various activities in which a person engages.

Whereas self-attributed motives are comparatively good predictors of immediate specific responses to structured situations, McClelland et al. continue, implicit motives are comparatively good predictors of long-term trends in behavior across various types of situations. Research findings described by McClelland et al. confirmed that indirect assessments of underlying motives have greater validity for predicting long-term trends in behavior than self-report assessments of motives that people directly attribute to themselves.

More recently, Bornstein (1995) has used a metaanalysis of 97 studies of measures of dependency to demonstrate further this difference between objective and projective assessment. With respect to differential prediction, according to Bornstein, available research indicates that objectively measured dependency correlates better with symptoms and the

diagnosis of dependent personality disorder than does projectively measured dependency, whereas projectively measured dependency correlates better with dependency-related behaviors.

The conceptual analysis formulated by McClelland et al. and elaborated by Bornstein has direct bearing on the contribution of projective methods to personality assessment. As previously described, projective assessment taps implicit motives and underlying personality characteristics that may not be readily apparent and may not be within a subject's conscious awareness. Because these covert motives and characteristics exert a powerful influence on long-term behavioral trends, this type of indirect measurement adds an important dimension to personality evaluations that would not be tapped in its absence.

Finally, with respect to what projective methods contribute to assessment batteries, the relative ambiguity of these methods makes them less subjective than structured instruments to influence by test-taking attitudes. This is not to say that projective methods are immune to subjects' efforts to present themselves in a positive or negative light. The relatively open-ended nature of projective testing situations and the dialogue they frequently elicit give subjects abundant opportunity to voice attitudes toward the tests, the examiner, and being examined. However, as long as subjects continue to give responses, neither their attitudes nor their expression of them is likely to prevent their projective test responses from revealing their personality characteristics. Simply put, the limited face validity of projective measures makes them more difficult to fake than objective measures, which means that they can balance a test battery to particularly good effect when self-presentation effects are of concern.

4.15.4.2 Empirical Basis

Projective measures vary in the extent to which they have been examined in well-designed research studies, and in many instances adequate empirical support for these instruments has lagged behind the uses to which clinicians sometimes put them. Nevertheless, the two most frequently used projective methods, the Rorschach and the TAT, have for the past generation been among the three most frequently studied personality assessment instruments, exceeded in this respect only by the MMPI/MMPI-2 (Butcher & Rouse, 1996). For both the Rorschach and the TAT, substantial evidence has accumulated to attest their validity for describing aspects of personality structure and dynamics and applying these descriptions in

differential diagnosis and treatment planning (Abraham, Lepisto, Lewis, & Schultz, 1994; Alvarado, 1994; Bornstein, 1995; Cramer & Blatt, 1990; Exner & Andronikoff-Sanglade, 1992; Ornduff & Kelsey, 1996; Ronan, Colavito, & Hammontree, 1993; Weiner, 1996; Weiner & Exner, 1991).

Later in this chapter, specific information is presented concerning the psychometric foundations and demonstrated corollaries of the projective measures most frequently used with young people. Suffice it to say in summary at this point that these measures prove valuable in personality assessment because they add information that would otherwise be unavailable and because they withstand relatively well efforts to exaggerate or conceal.

4.15.5 UTILITY OF PROJECTIVE ASSESSMENT

Whereas the value of projective techniques lies in the previously elaborated reasons *why* they should be used in personality assessment, their utility relates to decisions concerning *when* these methods should be included in a test battery. The kinds of information provided by projective test data indicate that projective measures should be used whenever a thorough personality assessment is considered relevant to formulating a differential psychodiagnosis or recommending alternative intervention strategies. Because of their relatively unstructured nature and indirect format, projective measures balance a test battery by tapping personality characteristics at a less conscious level than relatively structured measures. Assessments lacking such balance sample personality functioning from a limited perspective that will usually fail to paint a complete picture of the individual being examined. Batteries limited solely to projective techniques are similarly imbalanced and ill-advised in comprehensive personality assessments.

The previously mentioned conceptualization of McClelland et al. (1989) bears closely on the importance of a balanced test battery in clinical assessment. McClelland and his colleagues noted that measures of self-attributed and implicit motives seldom correlate with each other and should not be expected to do so, because they are measuring different aspects of personality. Moreover, given that directly and indirectly measured motives each predict certain kinds of behavior better than the other, they concluded that "Separate measures of self-attributed and implicit motives may be combined to yield a better understanding and prediction of certain types of behavior" (p. 692).

These formulations concerning different types of measures have subsequently been elaborated and confirmed for clinical purposes with respect to relationships between the Rorschach and the MMPI. Rorschach structural variables and MMPI scales have been found to show only a few modest correlations in both adult and adolescent samples (Archer & Krishnamurthy, 1993a, 1993b). At the same time, however, apparent contradictions between Rorschach and MMPI findings have been conceptualized by Weiner (1993, 1995b) not as invalidating either instrument or challenging the incremental utility of administering them in tandem, but rather as generative data.

Specifically, Weiner argues, apparently discrepant findings between personality assessment instruments of different kinds can be generative by virtue of complementing each other. Whereas findings on two tests that concur in suggesting the same personality characteristic are confirmatory and support definite conclusions, he continues, findings that diverge raise important questions to which they may also suggest helpful answers, especially if one of the tests is a relatively structured and the other a relatively unstructured instrument.

Consider, for example, a youngster who appears depressed on the Rorschach, with a high depression index (DEPI), but does not elevate on Scale 2 or the depression content scale of the MMPI-A. This divergent finding could well provide a useful clue to the adolescent's having an underlying or emerging depression that is not yet being keenly felt or manifest in well-structured situations. Alternatively, it could be that the subject is trying to deny or repress depressive affects and cognitions, or is making a conscious decision not to report manifestations of depression that nevertheless emerge in the absence of supportive structure or are revealed when the subject is uncertain how to conceal them.

The situation described in this example is familiar to assessment psychologists, who not infrequently work with psychologically troubled adolescents who can remain reasonably comfortable and controlled in relatively structured situations but become upset and disorganized in relatively unstructured situations and who may accordingly produce a benign MMPI-A protocol and a disturbed Rorschach. In such circumstances the objective measure has not erroneously overlooked psychopathology, nor has the projective measure mistakenly exaggerated it. Instead, the two types of test have combined in complementary fashion to provide valid information concerning the subject's likelihood of behaving in a relatively adaptive or maladaptive fashion,

depending on the situational context in which the behavior appears.

Similarly, subjects in some circumstances may produce a clinically unremarkable Rorschach while showing numerous elevations on the clinical and content scales of the MMPI-A. Such divergence is best understood not as error variance, but as a possible clue to the psychological stance of subjects whose degree of disturbance is minimal but who, when asked about themselves in language they can understand, want to make sure that others fully appreciate whatever problems and concerns they do have. Further illustrations of the clinical utility of divergence as well as convergence between a projective measure, such as the Rorschach and an objective measure such as the MMPI-2, are provided by Finn (1996) and Ganellen (1996).

4.15.6 APPLICABILITY OF PROJECTIVE METHODS TO CHILDREN AND ADOLESCENTS

Except for an occasional example, this chapter has thus far made no specific reference to young people. This apparent oversight is warranted by the fact that the nature of projective methods, the way in which they function, the kinds of information they provide, and the reasons for using them, are identical for persons of almost all ages. Hence the discussion of projective assessment to this point is as applicable to children and adolescents as to adults, and requires no modification or qualification as our focus now shifts specifically to young people.

Indeed, assessors who have learned to interpret projective test data provided by adults do not need to learn any new ways of working with the data should they begin to examine children and adolescents. By and large, the basic interpretive conclusions and hypotheses that attach to projective test variables apply regardless of the age of the subject. Whether they are age 8, 18, or 80, subjects who see numerous human figures on the Rorschach are likely to be quite interested in people; those who give long TAT stories are likely to be verbose; those who refer frequently to themselves in sentence completions are likely to be self-centered; and those who draw grotesquely distorted human figures probably harbor some disturbing concerns about their own nature or that of other people.

However, in order to determine the implications of these and other personality characteristics suggested by projective test data, examiners assessing young people must take

into account normative developmental expectations. For example, the data of developmental psychology indicate that children are more self-centered than adults, and subsequently become increasingly aware of and concerned about the needs of others as they grow through adolescence and approach maturity. Accordingly, test data that identify a high degree of self-centeredness may imply maladaptive narcissistic personality traits in an adult, but reflect normal development and adaptation in a child; conversely, minimal self-centeredness may indicate altruism and good adjustment in an adult but suggest deviant development and low self-esteem in a child.

Developmental psychology similarly provides some normative expectations for how children are likely to make drawings. Preschool age children commonly draw with what is called "intellectual realism," which means that they draw what they know to be there regardless of whether it would actually be visible. Thus, in x-ray fashion, young children often draw transparencies, such as people who are visible through walls (Di Leo, 1983). At about age 7 or 8, this intellectual realism gradually gives way to "visual realism," in which what is drawn resembles what realistically can be seen. Di Leo (1983, p. 38) observes that this developmental shift mirrors a metamorphosis in thinking from an egocentric to an increasingly objective view of the world. Hence a human figure drawing by a preschool child showing a belly button through clothing is much less likely to imply maladaptive functioning than the same drawing done by an adolescent.

As these examples indicate, familiarity with and adequate attention to normative expectation hold the key to valid and useful applications of projective methods in the assessment of young people. Ideally, projective methods manuals should include normative reference data that delineate quantitative as well as qualitative expectations for such developmental phenomena as maturational changes in self-centeredness. Regrettably, even though numerous projective test variables have been quantified in various ways, little progress has been made in generating age-graded norms for them.

The main exception to this dearth of normative developmental data for projective techniques is the Rorschach. Developmental trends in Rorschach responses from early childhood through adolescence were initially charted many years ago by Ames and her colleagues (Ames, Metraux, Rodell, & Walker, 1974; Ames Metraux, & Walker, 1971). More recently, the Rorschach comprehensive system has provided reference data for each of its codified variables on samples of approximately

100 nonpatient young people at each age from 5 to 16 (Exner & Weiner, 1995, chap. 3).

Aside from identifying needs for further research, an analysis of available data can guide clinicians in choosing which projective methods to include in a battery for assessing a young person's functioning. The more thorough and reliable the normative developmental data available for the instrument, the better the choice it will make. Similarly, the better established an instrument's correlates are in relation to behaviors that are central to the purpose of an assessment, the more reason there is to include it. Thus an instrument that has been demonstrated to be particularly helpful in identifying youthful depression may be a good choice in one case, whereas an instrument known to be especially sensitive in revealing family dynamics may be a good choice in another case.

Similarly, available empirical data and reported clinical experience should be drawn on to determine whether a particular instrument is likely to yield useful information concerning the personality functioning of individuals at certain ages. Thus the children's apperception test (CAT) depicting animal figures may be a more effective story-telling measure for a young child than the TAT, but certainly not for an adolescent (Bellak, 1993, p. 237).

As these observations indicate, projective methods provide sound clinical data only if they are employed in appropriate ways. First, examiners should have recourse to standardized procedures for administering and scoring any test they use. Lack of such standard methodology compromises the value of the data obtained, and inattention to standardized methods by examiners who opt instead for personalized approaches to administration and scoring is clinically disadvantageous and professionally questionable.

Second, clinical interpretations should be derived from test variables with demonstrated reliability and validity. Inadequate psychometric foundations limit the use to which test data can be put, and examiners who draw conclusions in the absence of supporting empirical evidence, without framing such conclusions as speculative hypotheses, are doing their patients and their methods a disservice. Third, the adequacy of projective assessment of young people will be limited in the absence of normative reference data for test responses of both adjusted and maladjusted children and adolescents and for developmental changes in these responses over time.

This chapter continues with reviews of the major inkblot, story telling, figure drawing, and sentence completion methods used in assessing

young people. The composition, administration, and scoring of each of these measures are described; what is known about their reliability, validity, and normative database is reported; and the clinical purposes they are likely to serve are discussed. The specific measures reviewed are selected primarily on the basis of their emphasis and frequency of use in clinical and school settings, as reported in surveys by Archer, Imhof, Maruish, and Piotrowski (1991), Elbert and Holden (1987), Hutton, Dubes, and Muir (1992), Kennedy, Faust, Willis, and Piotrowski (1994), Piotrowski and Keller (1989), Stinnett, Havey, and Oehler-Stinnett (1994), and Watkins, Campbell, Nieberding, and Hallmark (1995).

4.15.7 REVIEW OF PROJECTIVE ASSESSMENT METHODS

4.15.7.1 Rorschach Inkblot Method

The Rorschach inkblot method comprises 10 cards that are inked in shades of black and gray (five cards); black, gray, and red (two cards); and various pastel colors (three cards). The cards are reproduced in standard fashion, but the inkblot stimuli were originally designed at random and do not portray any specific objects (Rorschach, 1942). When subjects respond to the Rorschach, they draw on the shape, shading, and color of the blots to form impressions of what they might be, and in so doing they treat the instrument as a cognitive-perceptual task (e.g., "It looks like a bat, because it's got a body here and wings here and it's black"). In addition, subjects frequently elaborate their responses beyond the stimulus properties of the blots, and in so doing they treat the instrument as an associational task (e.g., "This bird is flying around looking for something to eat").

The cognitive-perceptual aspects of responses constitute structural data in Rorschach assessment and provide representative indications of the resources and coping style that a person generally brings to bear in problem-solving situations. The associational aspects of responses constitute thematic data in Rorschach assessment and provide symbolic clues to the underlying needs, attitudes, conflicts, and concerns that are likely to influence a person's actions and state of mind. The basic nature of the Rorschach in these respects is discussed further by Exner and Weiner (1995, chap. 1) and Weiner (1986, 1994).

4.15.7.1.1 Administration and scoring

The Rorschach is introduced by telling subjects that the inkblots they are about to

see are not anything in particular, but that people see many different things in them and that their task will be to indicate what the inkblots look like to them. The 10 cards are then given to subjects one at a time with the instruction "What might this be?" Requests for structure (e.g., "Can I turn the card?") are deflected back to the subject (e.g., "It's up to you"; "Any way you like"). The unguided responses to the 10 cards constitute the free association phase of the administration, following which there is an inquiry phase in which the examiner reads back each response and asks subjects where they saw it and what made it look as it did. The purpose of the inquiry is to facilitate coding of the structural features of the responses, and associations during this phase are not requested or encouraged. Responses are recorded verbatim, however, and the content of any spontaneous thematic elaborations is carefully noted.

Numerous approaches to codifying Rorschach responses have emerged during the long history of this instrument. However, for many years the comprehensive system of Exner (1993) has been by far the most widely used and researched (Piotrowski, 1996). Rorschach responses are coded in the comprehensive system for various aspects of where percepts are seen (location), why they look as they do (determinants), what they consist of (content), how commonly they occur (form level and populars), and whether they involve pairs of objects, organization of parts, or special kinds of elaborations, such as cooperative or aggressive interaction. These codes are then tallied and combined in various ways to yield a large number of indices, ratios, and percentages that guide the interpretive process, as elaborated in detail by Exner (1991, chaps 5–10).

4.15.7.1.2 Psychometric foundations

The psychometric foundations of an assessment instrument comprise the extent to which it can demonstrate adequate interscorer agreement, reliable measurement, valid correlates, and a representative normative database. The Rorschach inkblot method, as already indicated in part by examples used earlier in the chapter, rests on a solid psychometric basis. Interscorer agreement for the types of variables coded in the comprehensive system typically ranges from 80% to 100%. The reliability of Rorschach data has been demonstrated in a series of retest studies conducted over intervals ranging from seven days to three years and involving child, adolescent, and adult subjects. Most of the core variables associated with trait dimensions of personality show stability coefficients greater

than 0.80 in these studies, and some, including the affective ratio and the egocentricity index, consistently hover around 0.90 (Exner, 1991, pp. 459–460; Exner & Weiner, 1995, pp. 21–27; McDowell & Acklin, 1996; Weiner, 1997).

The validity of Rorschach assessment was confirmed in a series of metaanalytic studies that led Parker, Hanson, and Hunsley (1988) to conclude that the Rorschach meets usual psychometric standards for validity and is comparable to the MMPI in this respect. Specifically, Parker et al. used the effect sizes reported in 411 studies to derive population estimates of convergent validity of 0.41 for the Rorschach and 0.46 for the MMPI. Subsequent further confirmations of the validity of this instrument are noted by Weiner (1996).

With respect to its normative database, available information for the comprehensive system includes data on 700 nonpatient adults demographically representative of the 1980 US census, 1390 nonpatient children and adolescents age 5 to 16, and large samples of schizophrenic, depressed, and character disordered patients (Exner, 1993, chap. 12). In addition, longitudinal data reported by Exner, Thomas, and Mason (1985) on a group of young people tested every two years from age 8 to 16 provide useful reference information concerning developmental stability and change in Rorschach variables during childhood and adolescence.

As implied by the nature of the normative data, the Rorschach comprehensive system is applicable to young people from age five. Preschool age children have ordinarily not yet matured sufficiently to deal with the cognitive-perceptual aspects of the Rorschach situation in ways that lend themselves to the codification that is central to the comprehensive system interpretive process. An unusually mature four-year-old might on occasion produce a useful record, and immature five- and six-year-olds may produce records that have limited interpretive significance within the framework of the comprehensive system. Working within other frameworks, Ames et al. (1974) discuss and provide some normative findings for Rorschach responses of young children, and Leichtman (1996) has recently presented a developmental rationale for deriving information from the records of preschoolers.

4.15.7.1.3 Clinical utility

In common with projective techniques in general, the Rorschach serves clinical purposes primarily as a result of the information it provides about the structure and dynamics of personality functioning. With respect to

personality structure, the Rorschach has proved especially helpful in identifying and quantifying states of subjectively felt distress that combine elements of anxiety and depression and in reflecting trait dimensions of how people typically think, process information, handle emotions, manage stress, feel about themselves, and relate to others. Regarding personality dynamics, the thematic content of Rorschach responses, as previously noted, is often quite revealing of underlying needs, attitudes, conflicts, and concerns that influence how people are likely to think, feel, and act at particular points in time and in particular situations.

In addition, Rorschach data can frequently contribute to differential diagnosis in clinical settings. The comprehensive system provides indices for schizophrenia and depression (DEPI) that can help to identify these conditions in children and adolescents as well as in adults; for basic deficits in coping capacity that point to developmental arrest in young people; and for numerous features of conduct and anxiety and/or withdrawal disorders (Exner & Weiner, 1995, chaps 5–8; Weiner, 1986). Rorschach findings have also demonstrated considerable clinical utility in the treatment process by clarifying treatment targets, identifying potential obstacles to progress in therapy, and providing a basis for evaluating treatment change and outcome (Abraham et al., 1994; Weiner, 1994).

4.15.7.2 Thematic Apperception Test

The most widely known and used story telling technique is the TAT. It was developed by Morgan and Murray (1935) in the belief that the content of imagined stories would provide clues to the underlying dynamics of a subject's interpersonal relationships and self-attitudes. As elaborated by Murray (1943, 1971) and Bellak (1993, chap. 4), TAT data are expected to reveal the hierarchy of a person's needs and the nature of his or her dominant emotions and conflicts.

The TAT stimuli comprise 19 black-and-white illustrations of people or scenes and one blank card. The cards are intended for use with persons age five or older of both genders, and for nine of the cards there are alternate versions for use with adult and child/adolescent males and with adult and child/adolescent females. Because of the time required to administer the full set of TAT cards, examiners typically select a subset of 8–12 cards that they anticipate will elicit themes relevant to the assessment issues in a particular case. The themes usually elicited by the individual cards and the selection of subsets suited for children and adolescents are reviewed

by Bellak (1993, chap. 3), Dana (1985), and Obrzut and Boliek (1986). Regrettably with respect to standardization, however, there are no specific short forms of the instrument, and how many and which cards are typically chosen vary from one examiner to another and from one examination to the next.

4.15.7.3 Administration and scoring

The TAT cards are given to subjects one at a time with instructions to make up a story for each picture that includes (i) what is happening at the moment, (ii) what the characters are thinking and feeling, (iii) what led up to the situation, and (iv) what the outcome will be. The narrated stories are recorded verbatim by the examiner.

Murray (1943) originally proposed a scoring scheme in which each TAT story is rated for the presence and strength of a long list of needs that are being experienced by the central figure in the story and pressures that are being exerted by the environment. This scoring system proved too elaborate and time consuming for clinical work, and numerous alternative approaches to clinical interpretation were subsequently developed for the instrument. As reviewed by Chandler (1990), Murstein (1963), and Vane (1981), some of these interpretive approaches, like Murray's, have consisted of formal quantitative ratings of story characteristics. However, most interpretive approaches have eschewed quantitative scoring in favor of qualitative analyses of story content, and no scoring system has gained widespread use either clinically or in research studies.

The most commonly employed methods of interpreting the TAT in clinical practice appear to be variations of an "inspection technique" proposed by Bellak (1993, chap. 4). This technique consists simply of reading through subjects' stories to identify repetitive themes and recurring elements that appear to fall together in meaningful ways. Because this approach lacks any quantification and rests on the capacity of individual examiners to relate story themes and elements to aspects of personality functioning, Dana (1985) was moved to observe that "TAT interpretation has become a clinical art form" (p. 90).

Bellak's influential approach stresses 10 aspects of a story, each of which is taken to have implications for how subjects view and are likely to deal with interpersonal events and what they anticipate the future to hold for them. These include the main theme of the story, the identity of the central figure or hero, the main needs of the hero, the way in which the

environment is conceived, the identity and intentions of other figures in the story, the nature of any anxiety or other affect that is being experienced, the nature of any conflict that is described or suggested, the ways in which conflicts and fears are defended against, the ways in which misbehavior is punished, and the level of ego integration.

With respect to research studies, the most productive utilization of the TAT has derived from quantitative scoring systems developed by McClelland, Atkinson, and their colleagues to measure needs for achievement, affiliation, and power (Atkinson & Feather, 1966; McClelland, Atkinson, Clark, & Lowell, 1953). Although scoring for achievement, affiliation, and power motivation has had little clinical impact, other schemes for coding specific personality characteristics reflected in TAT thematic content have subsequently emerged. These include scales for level of ego development (Sutton & Swenson, 1983), preferred defense mechanisms (Cramer, 1987), quality of interpersonal affect (Thomas & Dudek, 1985), problem-solving style (Ronan et al., 1993), and object relations capacities (Westen, Lohr, Silk, Kerber, & Goodrich, 1985).

Particularly promising among these is the use of the TAT to assess aspects of object relatedness through the Westen et al. (1985) measure, known as the social cognition and object relations scale (SCORS). The SCORS provides quantitative indices of the affective tone subjects ascribe to relationships, their capacity for emotional investment in relationships and social standards, their understanding of social causality, and the complexity of their representations of people. By including ratings of subjects along dimensions of maturity as well as normality/pathology, the SCORS is proving especially relevant to the assessment of young people (Westen et al., 1991).

4.15.7.3.1 Psychometric foundations

Efforts to demonstrate the reliability and validity of global approaches to interpreting the TAT have been handicapped by the previously noted proliferation of scoring systems, by clinicians' preferences for a strictly qualitative and uncoded approach to the data, and by enormous variation in how the test is administered, including which subset of cards is selected for use. Because of this long-standing lack of standardization, there has been little opportunity for systematic accumulation of data bearing on the reliability and validity of the TAT in general, nor has it been possible to develop any substantial normative database.

Accordingly, for both inspection techniques and overall scoring systems developed in the tradition of Murray, the psychometric literature on the TAT is generally acknowledged to comprise a mix of positive and negative findings that cannot easily be compared with one another. Hence, despite the widespread use of inspection techniques in clinical practice, neither these nor other global approaches have been demonstrated to show adequate psychometric properties.

However, research studies with TAT scales developed to measure specific personality characteristics have demonstrated that the instrument can generate reliable and valid findings when it is used in a standardized manner. The previously mentioned scales of Cramer (1987) and Westen et al. (1985) are cases in point. Cramer's scale reliably identifies three major mechanisms of defense—denial, projection, and identification—and has shown valid corollaries in changes observed in patients undergoing psychotherapy (Cramer & Blatt, 1990). The Westen et al. SCORS has been found to provide reliable identification of developmental variables related to disturbed object relations in children and, as already mentioned, is therefore especially relevant to the assessment of young people. Validation studies with SCORS have involved psychiatrically disturbed, borderline, physically abused, and sexually abused young people (Freedefeld, Ornduff, & Kelsey, 1995; Ornduff, Freedefeld, Kelsey, & Critelli, 1994; Westen, Ludolph, Block, Wixom, & Wiss, 1990; Westen, Ludolph, Lerner, Ruffins, & Wiss, 1990).

4.15.7.3.2 Clinical utility

The clinical utility of the TAT lies mainly in its potential for elucidating dynamic aspects of personality functioning, particularly with respect to the feelings and attitudes that subjects hold towards other people, themselves, and possible turns of fortune in their lives for better or worse. Based on the assumption that children and adolescents identify with the central figures in their TAT stories and project fantasies and realities regarding their own lives into the events and circumstances they describe, the obtained data can shed light on a broad range of underlying influences on how young people are likely to think, feel, and act.

As previously noted in commenting on the research of McClelland et al. (1989), the implicit types of motives measured by the TAT are more likely to correlate with persistent dispositions to behave in certain ways rather than with immediate actions or symptom formation.

Accordingly, TAT findings will usually not add very much to structural diagnosis of adjustment problems in young people, but they can be extremely helpful in suggesting possible dynamic origins of adjustment problems. In this regard, the psychometrically sound SCORS may be a useful scale to include in forensic assessment batteries when issues of custody or adoption are being addressed. This TAT scale can frequently assist examiners in grasping a young person's representations of people and his or her capacities for emotional investment in relationships.

The thematic content of TAT stories has additional potential to facilitate planning and conducting psychotherapy with young people, particularly with respect to identifying treatment targets and monitoring progress in therapy. The TAT can also be used in treatment as a play therapy tool, as in Gardner's (1971) story telling technique. For example, after a youngster has told TAT stories, the therapist and child can act out the stories in play, or the therapist can create stories to the same picture stimuli for comparisons with the child's stories. Hoffman and Kupperman (1990) describe such an intervention with a 13-year-old boy in which both therapists wrote stories to the same TAT cards to which the patient had responded. As it turned out, Hoffman's stories emphasized the main character's maladaptive coping mechanisms, whereas Kupperman's stories emphasized positive and healthy aspects of the central character's coping capacities. Over a number of sessions, this boy and his therapist engaged in discussions concerning whose version of the story was most accurate.

4.15.7.4 Children's Apperception Test

Consistent with the purpose of the TAT, Bellak (1993, chap. 13) developed the CAT to facilitate understanding of personality processes in children, including their "dynamic way of reacting to and handling the problems of growth" (Bellak & Siegel, 1989, p. 102). The CAT pictures were designed to elicit fantasies about aggression, sibling rivalry, fears of being alone at night, attitudes toward parental figures, and eating problems.

The CAT-Animal (CAT-A) form, originally published in 1949 and designed for children 3–10 years old, consists of 10 pictures depicting animals in human situations. The use of animal figures was based on the assumption that young children identify more readily with animals than with people and will accordingly tell more meaningful stories about animal than human figures. Moreover, according to Bellak (1993,

chap. 13), the use of animal figures makes the CAT-A a culture-free test that is equally applicable to Caucasian, African-American, and other minority group youngsters as well as to children from different countries, except where there is little familiarity with some of the inanimate objects depicted, such as bicycles.

There is also a human form of the CAT (the CAT-H) that was developed by Bellak and Hurvich (1966) in response to criticism of the assumption that children identify more easily with animal than with human figures. Studies reviewed by Bellak and Hurvich indicate little difference in stimulus value between the original CAT-A and the CAT-H, in which human figures are substituted for the animals in the CAT-A scenes. However, the CAT-H does not appear ever to have become much used in clinical practice.

4.15.7.4.1 Administration and scoring

Children being administered the CAT are told that they are going to take part in a game in which they will tell stories about pictures. Subjects who appear to regard the CAT as a test are informed that it is not the type of test in which they will be graded for correct or incorrect answers. Standard procedures call for all 10 CAT pictures to be administered in numerical order, from Card 1 to Card 10. Children are told to narrate what the animals are doing in the pictured scenes and are asked at appropriate points to say what went on previously and what will happen next. The examiner encourages and prompts the subject as necessary but avoids being suggestive or asking leading questions. Examiners may also query each story by asking the child to elaborate specific points such as the ages of characters and why they were given particular names.

In clinical work the CAT is typically interpreted along the lines proposed by Bellak for the TAT, that is, with an inspection technique used to form qualitative impressions of various dimensions of the subject's personality functioning (Bellak, 1993, chap. 14). There is an alternative but rarely used quantitative approach developed by Haworth (1965), called the schedule of adaptive mechanisms, in which CAT responses are rated numerically for the degree of adaptability or disturbance they reflect. Haworth also stressed the importance of recognizing that young children are highly reactive to the immediate circumstances in their lives and less likely than adolescents or adults to have formed well-established personality traits. She accordingly emphasized careful interpretation of CAT responses in the context of adequate information concerning subjects'

home situation and the nature of any recent or impending crises in their lives.

4.15.7.4.2 Psychometric foundations

There has regrettably been little accumulation of empirical data bearing on the reliability and validity of the CAT. The widespread use of Bellak's qualitative inspection technique in CAT interpretation and a corresponding lack of quantification have precluded examination of the instrument's psychometric foundations. Although it has sometimes been suggested that the idiographic nature of CAT as well as TAT data makes traditional psychometric criteria difficult to apply or even irrelevant, there is nothing in the nature of the data generated by story telling techniques that prevents their being reliably coded for various types of feelings, motives, attitudes, and capacities that can in turn be validated against meaningful correlates. The previously noted development of psychometrically sound TAT scales for such specific aspects of personality as achievement motivation and social cognition proves the point that clinical interpretation of stories can go beyond being an art form and attain respectability as a scientific procedure as well.

Regarding what research is available concerning the CAT, Bellak (1993, chap. 16) provides a review of studies comparing the responses typical of children at different ages and examining special features in the stories of maladjusted, schizophrenic, speech disordered, retarded, brain-damaged, and chronically ill children. Almost all of these studies date from the 1950s, however, and none provides an adequate basis for developing any formal normative standards or diagnostic guidelines for the instrument.

4.15.7.4.3 Clinical utility

As in the case of the TAT, the CAT is useful in clinical assessment primarily as a source of hypotheses concerning subjects' personality dynamics, particularly with respect to how they view themselves and other important people in their lives, the nature of their hopes and fears, and what they expect will happen to them. Although not suitable for adolescents, the CAT is more useful than the TAT in work with younger children, who are likely to relate more easily to the familiar situations and youthful figures depicted in the CAT illustrations than to the primarily adult figures and unpopulated scenes shown in the TAT.

Like the TAT, the CAT may also contribute to treatment planning, by suggesting areas of concern on which to focus in the therapy, and it

may itself serve as a play technique. In diagnostic assessments, however, hypotheses generated by this instrument require support from other data prior to being addressed to questions that necessitate empirical decision making.

4.15.7.5 Roberts Apperception Test for Children

The Roberts apperception test for children (RATC) is intended for use with young people of ages 6–15 and was designed to improve on the TAT and CAT by presenting familiar stimuli and employing a standardized scoring system (McArthur & Roberts, 1990).

Instead of illustrations primarily of adults, such as those used in the TAT, or illustrations of animals, as used in the CAT, the RATC primarily portrays child and adolescent figures engaged in everyday interactions, including scenes of parental affection, disagreement, school and peer relationships, and observation of nudity. There are 27 RATC cards, 11 of which are alternate versions for male or female subjects. There is an addition an alternate set of cards portraying African-American individuals in similar scenes.

4.15.7.5.1 Administration and scoring

The standard 16 cards that compose the RATC are administered individually in numerical order, using male or female versions as appropriate. Subjects are instructed to make up a story about each picture and to tell what is happening in the picture, what led up to the scene, how the story ends, and what the people are talking about and feeling. Responses are recorded verbatim.

If subjects tell an incomplete story or omit certain aspects, such as how the characters are feeling, additional inquiry may be used to help teach them to give scorable responses. This inquiry may be used liberally with the first two cards but not thereafter, which means that responses on cards 3 through 16 may at times have limited scorable data. The possibility of limited scorable data is the price to be paid for maintaining careful standardization of the RATC procedure. McArthur and Roberts sanction deviations from the standard procedures only in specific instances, such as the examination of severely disturbed children. When standard procedures are not or cannot be followed, they recommend considerable caution in using the scoring procedures.

The scoring system for the RATC consists mainly of coding the content of each story for the presence or absence of 13 profile scales that

have implications for personality characteristics. Eight of these are adaptive scales that relate to thematic indications of reliance on others, giving support to others, supporting oneself, limit-setting by authority figures, identification of problem situations, and resolving problems in unrealistic, constructive, or particularly insightful ways. The other five are clinical scales that pertain to thematic manifestations of anxiety, aggression, depression, experiences of rejection, and inability to resolve problems. The test manual provides guidelines and examples that promote reliable scoring of each of these scales.

To facilitate interpretation of the data, the raw scores for each of the 13 profile dimensions are summed over the 16 cards and then plotted on a normatively scaled profile form to yield a visual representation of the data, much in the manner of an MMPI-A profile. The interpretive yield of the data is further enriched by use of an interpersonal matrix, which consists of a tabular representation of the frequency of convergence between particular scales and the figures identified (e.g., cooccurrence of themes of reliance on others with description of interaction with a maternal figure).

4.15.7.5.2 Psychometric foundations

As with the careful coding of specific variables on the TAT, the development of standardized administration and scoring procedures for the RATC has demonstrated that story telling projective methods can achieve psychometric respectability. McArthur and Roberts (1990) report inter-rater agreement ranging from 0.80 to 0.93 in the scoring of the various dimensions in their system and split-half reliabilities ranging from 0.44 to 0.86 on their profile scales, with half of these scales showing reliability coefficients of 0.73 or higher. The adaptive scales appear to work better in this regard than the clinical scales: only one of the five clinical scales (inability to resolve problems) shows a split-half reliability greater than 0.55, whereas all but one of the eight adaptive scales (reliance on others) shows a split-half reliability above 0.60.

Validity studies conducted with 200 non-patient and 200 outpatient youngsters aged 6–15 indicate that subjects generally tell stories that fit the expectations for the cards for example, that a card intended to depict parental affection typically elicits stories involving affection. In addition to thus showing content validity, the RATC profiles have been found to distinguish between well-adjusted and clinic youngsters, to be more resistant to efforts at manipulation than self-report measures, and to correlate well

with a behavior problem checklist completed by subjects' parents (McArthur & Roberts, 1990; Worchel, Rae, Olson, & Crowley, 1992). Significantly, however, in the comparisons between well-adjusted and clinic youngsters, significant differences were found on all eight of the adaptive scales and on two of the five clinical scales, but not on the clinical scales for anxiety, aggression, and depression.

With respect to its normative database, the RATC manual provides the mean and standard deviation for each of the profile scales for 200 nonpatient youngsters divided by gender and by four age categories (6–7, 8–9, 10–12, and 13–15).

4.15.7.5.3 Clinical utility

Similar to the TAT, CAT, and other story-telling measures, the RATC contributes to personality assessment of young people primarily by casting light on their underlying attitudes and concerns. Although the measure includes specifically designated clinical scales that might be expected to assist in clinical diagnosis as well as dynamic analysis of a youngster's personality functioning, its clinical scales as noted appear less reliable than its adaptive scales and less capable of discriminating between patient and nonpatient populations.

The RATC is especially valuable to clinicians as well as researchers by virtue of its careful standardization and adequate psychometric properties. Aside from facilitating the interpretive process and providing the basis for systematic accumulation of data, these features of the instrument make it particularly attractive to examiners conducting forensic evaluations. Having used the RATC in their projective assessment of a child or adolescent, rather than the TAT or CAT, psychologists will be better prepared to justify their procedures and conclusions when giving testimony as an expert witness.

However, examiners may sometimes question whether the information they get from the RATC warrants the amount of time required to administer it, which in our experience approximates an hour on the average. For some young children, moreover, there may be difficulties in sustaining their investment in the task. There is no short form of the RATC to use and no shortcut, such as selecting just a few cards to use or discontinuing a protracted administration before giving all 16 cards. Doing so eliminates the standardization of the instrument and prevents any meaningful scoring or comparisons with normative data, which means that the test loses its advantage over a TAT or CAT interpreted by inspection. Our view in this matter is that the RATC warrants the time

required to administer and score it in the standardized manner and, more generally, that psychologists should if at all possible take whatever time is necessary to use standardized instruments when they are available.

The alternate form of the RATC involving African-American figures may prove useful in assessing African-American young people. However, there are as yet no data establishing the reliability and validity of this alternate form, nor have any representative normative data been published for it. Hence the RATC should be used cautiously in multicultural assessments, for which at present the most promising instrument is Tell-me-a-story.

4.15.7.6 Tell-me-a-story

The Tell-me-a-story (TEMAS) was designed as a multicultural story-telling test based on the concept that personality development occurs within a sociocultural system in which individuals internalize the cultural values of their family and society (Costantino, Malgady, & Rogler, 1988). The TEMAS is intended for use with African-American, Hispanic, and Caucasian children and adolescents aged 5–18 and comes in two parallel sets, one for minority and the other for nonminority youngsters. The two sets feature either predominantly Hispanic and African-American characters or predominantly nonminority characters all shown in urban environments. Each set comprises 23 cards, 11 of which have alternate versions for males and females and one of which has alternate versions for children and adolescents.

As a distinctive feature of the TEMAS, many of the pictures portray a split scene showing contrasting or conflicting intrapersonal and interpersonal situations that require some resolution, much in the manner of Kohlberg's (1976) moral dilemma stories. For example, one side of a scene may depict apparent delay of gratification, and the other side an inability to delay gratification. How subjects resolve this conflict in their stories speaks to the adaptiveness of their personality functioning and their stage of moral development.

The TEMAS provides quantitative scales for measuring the adequacy of subjects' adaptation with respect to nine aspects of personality functioning: interpersonal relations, aggression, anxiety/depression, achievement motivation, delay of gratification, self-concept, sexual identity, moral judgment, and reality testing. There are also quantitative scales for four cognitive functions related to how individual process information (reaction time, total time, fluency as reflected in the number of words used

in a story, and omissions of relevant visual details) and for four affective functions as indicated by mood states attributed to the main characters in a story (happy, sad, angry, and fearful). These quantitative scales are supplemented by several qualitative indicators used to describe various other characteristics of the stories.

As stressed by Costantino, Malgady, and Rogler (1988), the TEMAS was developed to overcome limitations of traditional thematic apperception tests and differs from the TAT in several significant ways. These include a focus on interpersonal relationships rather than on intrapsychic dynamics; the use of personally relevant and culturally sensitive stimuli that emphasize meaning rather than ambiguity; the representation of both positive and negative poles of emotions, cognitions, and interpersonal functions, as opposed to the heavy weighting of the TAT stimuli with representations of depression, gloom, anger, and hostility; and the introduction of the joint depiction of contrasting circumstances to elicit expressions of conflict resolution and moral judgment.

4.15.7.6.1 Administration and scoring

Subjects are administered either the minority or nonminority form of the TEMAS, and examiners can choose between a long form, which comprises all 23 cards and requires approximately two hours to administer, and a standard short form, which consists of nine cards and requires one hour to give. In keeping with generally recommended practice in multicultural assessment (Dana, 1993, chap. 6, 1996), young people should be tested in their primary language, and those who are bilingual should be tested by a similarly bilingual examiner. Subjects are instructed to tell a complete story about each picture that indicates what is happening now, what happened before, and what will happen in the future. These instructions may be repeated as often as necessary and supplemented with structured inquiries to elicit information concerning who and where the characters are and what they are thinking and feeling.

The TEMAS is scored by rating stories for the presence of the various cognitive and affective functions and on a four-point scale for the level of adjustment reflected in thematic indications of the personality functions. The resulting scores are then totaled and translated into normalized T scores that are graphed to provide a readily interpretable visual profile. The TEMAS manual provides detailed guidelines and case examples to illustrate these scoring procedures.

4.15.7.6.2 Psychometric foundations

As described in the test manual (Costantino, Malgady, & Rogler, 1988), the TEMAS was carefully developed and standardized over several years prior to its publication. Unfortunately, except for continued work by the test's authors, much of it in the form of paper presentations, there has been little published research concerning the psychometric adequacy of the instrument. On balance, however, the data provided in the test manual appear to indicate a promising beginning in demonstrating its soundness.

With respect to inter-rater agreement, studies with Hispanic, African-American, and Caucasian subjects have indicated that both the minority and nonminority versions of TEMAS can be scored reliably, with agreements between trained examiners generally ranging from 75% to 95% across the various scales. Regarding issues of reliability, on the long form of the TEMAS 11 of the 17 personality, cognitive, and affective scales showed internal consistency (alpha) correlations of 0.74 or higher in the standardization data, and the median for all 17 was 0.83. Internal consistency was lower on the short form, with a median of 0.68. The internal consistency data for the short form must be considered preliminary, however, because in this analysis the short form scores were extracted from the protocols of subjects who had in fact completed the long form, rather than from actual administration of the short form. In an assessment of its retest reliability, the short form was administered twice to 51 behavior problem children over an 18 week interval. Very little stability over time was demonstrated in this study. Costantino, Malgady, and Rogler (1988) suggest several plausible explanations for this disappointing result, such as the narrow range of scores among the subjects. Nevertheless, test-retest reliability remains to be demonstrated for both the short and long forms of the instrument.

Turning to the validity of the TEMAS, ratings by psychologists of the types of personality functions pulled by the stimulus pictures have indicated good agreement with the intent in designing them, thus attesting that the test measures what it purports to measure. In terms of its criterion validity, the TEMAS has been found in several studies to discriminate between patient and nonpatient youngsters in Hispanic and African-American as well as Caucasian samples, and to accomplish this distinction in inner city as well as middle-class settings (Costantino, Malgady, Bailey, & Colon-Malgady, 1989; Costantino, Malgady, Colon-Malgady, & Bailey, 1992; Costantino,

Malgady, Rogler, & Tsui, 1988). The TEMAS has not demonstrated any capacity to differentiate specific kinds of disorders, but there is some evidence to suggest that certain story characteristics, including the omission of main characters or events and failure to notice a conflict in the picture (due to lack of attention), may be sensitive to the presence of attention deficit hyperactivity disorder (Costantino, Colon-Malgady, Malgady, & Perez, 1991).

Initial studies reported in the TEMAS manual indicate further that many of its scales correlate significantly with behavior ratings by subjects' mothers and teachers and with behavioral observations of their inclinations toward aggressive or disruptive behavior and their capacities for self-confidence and delay of gratification. Preliminary data indicate further that TEMAS profiles of young people prior to their entering therapy can significantly predict aspects of their treatment outcome.

The TEMAS standardization sample comprised 281 male and 361 female youngsters from the New York City area who ranged in age from 5 to 13 years. This sample of 642 children and adolescents included groups of Caucasian, African-American, Puerto Rican, and other Hispanic subjects from predominantly lower- and middle-income families. The test manual lists the mean and standard deviation for each scale by gender and ethnicity for the age groups 5–7, 8–10, and 11–13.

4.15.7.6.3 *Clinical utility*

The TEMAS brings to clinical assessments the distinct advantages of a well conceptualized and quantitatively standardized story-telling technique that is also culturally sensitive and normed for minority groups of young people. As in the case of the RATC, its coded scales and profile graphs facilitate interpretation and provide the type of documentation that typically proves valuable in forensic cases. Moreover, this measure stands alone as a story-telling test proved applicable in the assessment of African-American and Hispanic children and adolescents. The TEMAS accordingly merits serious consideration for inclusion in a test battery for evaluating personality functioning in young people, especially if they come from an urban minority background and if there are forensic issues in the case.

At this stage in its development, however, the TEMAS has some drawbacks that examiners should keep in mind. First, although it appears to be useful for assessing level of adjustment and incorporation of societal norms among both minority and nonminority youth, TEMAS does

not provide the breadth of information concerning specific types of concerns and relationships that emerge from TAT, CAT, and RATC analyses. Second, the normative data available thus far are limited to 5–13-year-old city dwellers and thus do not provide a psychometric basis for drawing conclusions about adolescents aged 14–18, or suburban and rural dwelling youngsters. Third, the length of time (two hours) required to administer the 23 card long-form of the test is often impractical in clinical evaluations. The nine card short-form is an attractive alternative but, as noted, adequate reliability has not yet been demonstrated for the short-form.

4.15.7.7 *Draw-a-person*

The use of human figure drawings as a projective method of personality assessment is based on the expectation that how subjects draw people will reveal aspects of how they perceive themselves and feel about others. Clinical use of drawing techniques originated in work with children, and the first formal drawing test, called the draw-a-man, was developed by Goodenough (1926) and later refined by Harris (1963) and Naglieri (1988) for use as a nonverbal measure of intellectual development. Machover (1948) introduced the notion of using human figure drawings as a projective device that generates nonverbal, symbolic messages concerning subjects' impulses, anxieties, and conflicts, and she proposed numerous possible meanings for structural features of drawings (e.g., where figures are placed on the page) and the manner in which various parts of the body are drawn (e.g., a disproportionately large head).

Koppitz (1968, 1984) subsequently focused attention specifically on draw-a-person (DAP) assessment of young people and used on Machover's interpretive hypotheses to formulate 30 specific indicators of emotional disturbance involving the quality of drawings (e.g., asymmetry, transparencies), special features (e.g., teeth showing, arms clinging to body), and omission of body parts (e.g., no eyes, no arms). Naglieri and his colleagues (Naglieri, McNeish, & Bardos, 1991; Naglieri & Pfeiffer, 1992) have developed further coding refinements to produce the draw-a-person: screening procedure of emotional disturbance (DAP:SPED). The DAP:SPED is an actuarially derived and normatively based system comprising 55 objectively scorable items, such as the measured dimensions and placement of figures. It is intended as a screening test for classifying young people aged 6–17 with respect to their

likelihood of having adjustment difficulties that call for further evaluation.

4.15.7.7.1 Administration and scoring

The DAP is administered by giving subjects a plain 8.5×11 inch piece of paper and asking them to draw a person. When they have finished the drawing, they are given another piece of paper and asked to draw a person of the opposite sex from the one they have just drawn. Subjects are further instructed to draw the figure of a whole person rather than a cartoon or stick figure. In keeping with further suggestions by Machover (1951), subjects are also typically asked to draw a picture of themselves and to provide some thematic content concerning the figures they have drawn.

As elaborated by Handler (1985), there are three alternative ways of eliciting such thematic content: by asking subjects to associate to their drawings, by asking them to make up a story about the people they have drawn, or by asking them specific questions about their drawings. Machover (1951) provided 31 specific questions to be used for this purpose with children, such as "What is their ambition?," "How happy are they?," "What do they worry about?," and "What are their good points?"

As with story-telling techniques, figure drawing methods are most commonly interpreted in clinical practice by an inspection in which personality characteristics are inferred primarily from subjective impressions of noteworthy or unusual features of the figures drawn. Items included in the Koppitz and DAP:SPED scales may often enter into these impressionistic assessments. However, these scales serve only to identify emotional disturbance without contributing in other ways to personality description, such as by indicating how individuals process information, handle emotion, and manage stress. They have consequently not been widely adopted clinically, and there are no other DAP scoring systems that have attracted much attention in the research literature.

4.15.7.7.2 Psychometric foundations

There is very little psychometric foundation for traditional applications of the DAP, and clinical use of this instrument in assessing young people frequently goes well beyond any empirical justification in empirical data. The influential inspection approach used by Machover is neither standardized nor codified, which precludes any systematic evaluation of its reliability or the accumulation of a normative database. Additionally, although it may be reasonable to

expect that unusual emphasis on or omission of some body part will reflect some particular concern about the nature or functions of that body part, most of Machover's specific hypotheses concerning the symbolic significance of figure drawing characteristics lack consistent research support (Kahill, 1984; Roback, 1968; Swensen, 1957, 1968).

The Koppitz system is adequately codified and intended to comprise items that have demonstrably low occurrence in the normal population. However, there is no normative database for the system, its reliability is yet to be demonstrated, and there is some question as to whether it can differentiate between well-adjusted and emotionally disturbed children. In a carefully done study in which Tharinger and Stark (1990) compared groups of mood disordered, anxiety disordered, mood/anxiety disordered, and well-adjusted youngsters aged 9.5–14.75 years, the Koppitz signs showed good interscorer agreement but did not differentiate among the subject groups either in mean total score or in the frequency of any of the 30 individual items.

Such findings warrant concern that the DAP may be a faulty projective technique with questionable propriety for continued clinical use. However, it could be that the DAP is a potentially sound method for which there has not yet been sufficiently sophisticated development and evaluation to document its capacities. The previously mentioned work of Naglieri et al. (1991) on the DAP:SPED appears to speak to this point. The DAP:SPED was standardized on a representative national sample of 2355 children and adolescents aged 6–17 years; its objective scoring procedures have generated inter-rater agreements above 90%; its internal consistency (alpha) reliability estimates are 0.76 among 6–8-year-olds, 0.77 among 9–12-year-olds, and 0.71 among 13–17-year-olds; and its normalized total score has shown substantial capacity to differentiate nonpatient youngsters from those with identified behavioral or emotional problems (McNeish & Naglieri, 1991; Naglieri & Pfeiffer, 1992).

Also noteworthy is the work of Tharinger and Stark (1990), who paired their failure to validate the Koppitz system with an investigation of a proposed new system of their own, the DAP integrative system. The integrative DAP is based on a qualitative holistic scoring approach in which drawings are given an overall adjustment rating on a scale from 1 (absence of psychopathology) to 5 (severe psychopathology). Examiners are instructed to base their impressions on their integrated sense of four characteristics of a drawing, with the pathological end of the scale involving (i) inhumanness

of the drawing suggesting feelings of being incomplete, grotesque, or monstrous; (ii) lack of agency as conveyed by a sense of powerlessness; (iii) lack of well-being as reflected in negative facial expressions; and (iv) a hollow, vacant, or stilted sense indicating lack of capacity to interact.

In the same study in which Tharinger and Stark failed to validate the Koppitz system, their integrative system total score significantly discriminated the mood disordered and anxiety/mood disordered subjects from the well-adjusted youngsters and also correlated significantly with the Coopersmith self-esteem inventory, thus attesting the capacity of the DAP to depict a youngster's sense of self. Contemporary literature abounds with sharply divided opinion concerning whether the DAP is a worthless test that should no longer be used (e.g., Gresham, 1993; Motta, Little, & Tobin, 1993) or is instead a potentially valuable clinical tool that has too often been carelessly used or inadequately researched (e.g., Bardos, 1993; Holtzman, 1993). Results to date with the DAP:SPED and the DAP integrative system give some reason to believe that improved methodology may yet establish sound psychometric foundations for carefully specified applications of the DAP.

4.15.7.7.3 Clinical utility

Despite its widespread use in describing subjects' personality characteristics, there is virtually no empirical evidence that traditional DAP interpretation has any clinical utility. Smith and Dumont (1995) asked a group of experienced psychologists and graduate students who had been trained in the use of the DAP to review and comment on a case file. They found that these clinicians routinely utilized specific symbolic representations in the drawings to draw inferences about the client's personality characteristics and diagnostic status—even though research does not support any such isomorphic correspondence of specific drawing characteristics to specific features of personality. Aside from the general ethical issues of using test instruments in unwarranted ways, examiners who are preparing forensic testimony jeopardize their credibility by employing the DAP in this manner.

This does not mean that human figure drawings are without utility in the clinical assessment of children and adolescents. First, by contrast with the practical difficulties of administering the full RATC and TEMAS, the DAP is an easily administered measure that requires no test stimuli and usually takes less than 10 minutes.

Second, as a nonverbal measure the DAP can prove especially useful in the evaluation of frightened, reticent, or otherwise uncommunicative young people, and it is unaffected by language difficulties or bilingualism. As noted by Cummings (1986), moreover, many young children may be more capable of expressing their thoughts and feelings in drawings than in words, and drawing pictures is a more familiar activity to most youngsters than most of the tasks that are set for them in a psychological examination.

Third, unusual characteristics of drawings can suggest avenues for further exploration, even in the absence of definitive conclusions about what these features signify. Examiners need to be circumspect in pursuing such avenues, however, lest they prematurely conclude that they are exploring in correct directions. The pursuit of speculative hypotheses is just as likely to lead into blind alleys as down fruitful paths.

Finally, there is ample indication in the contemporary literature that such approaches as the DAP:SPED and the DAP Integrative System can help to identify the presence and severity of emotional disturbance in general and can accordingly contribute to treatment recommendations, planning, and monitoring. In the future, codifiable and interpretable thematic content in subjects' stories about their drawings, comparable to the data derived from story telling techniques, may further enhance the utility of the DAP.

4.15.7.8 House-tree-person

The House-tree-person (HTP) test was devised by Buck (1948, 1985) as a means of tapping the concerns, interpersonal attitudes, and self-perceptions of young people more fully than is possible with human figure drawings alone. The HTP is intended for use with anyone over the age of three and was regarded by Buck as a nonthreatening instrument that can serve well to minimize a child's anxiety in testing situations and to assess personality functioning in multicultural and bilingual settings. As postulated by Buck and subsequently elaborated by Hammer (1958, 1985), subjects' drawings of these three objects are considered to provide symbolic representations of important aspects of their world.

Specifically, the house, as a dwelling place, is expected to arouse feelings toward the subject's home life and family relationships and, particularly for children, attitudes toward their parents and siblings. The tree is seen as encouraging projection of personal feelings

about the self that would be more anxiety-provoking to express in drawing a person, because the latter is more obvious in its representation as a self-portrait. More specifically, the way the trunk of the tree is drawn is considered to portray a subject's feeling of basic power and inner strength; the branches are seen as depicting the subject's ability to derive satisfaction from the environment; and the overall organization of the drawing is taken as a reflection of the individual's feeling of intrapersonal balance. Finally, the drawing of the person is expected to reveal aspects of how subjects view themselves, how they would like to be, and what they think about significant other people in their lives.

4.15.7.8.1 Administration and scoring

Although Buck recommends a four-page booklet with pages measuring 7×8.5 inches, most HTP examiners use four sheets of standard 8.5×11 in paper on which subjects are asked first to draw "as good a picture of a house as you can," then a tree, and then a person of each sex. Subjects are told they can take as long as they wish and draw any kind of house, tree, or person they like. Completion of the drawings is followed by an interrogation phase in which numerous questions devised by Buck are used to encourage subjects to define, describe, and associate to their drawings (e.g., "About how old is that tree?," "What does that house make you think of?" "Is that person happy?") Buck also recommended a chromatic phase of the HTP in which subjects would do a second rendering of their drawings in crayon rather than pencil, followed by another interrogation. There are no data to indicate the frequency with which examiners conduct Buck's full HTP administration or instead limit the test to the pencil drawings, without either an inquiry or a chromatic phase.

Following the example of Goodenough, Buck originally proposed an elaborate quantitative system for objective coding of structural features of the HTP drawings, such as their size and proportions. As best as can be determined from the literature, however, quantitative coding of the HTP has rarely been employed in clinical practice. Instead, clinicians using this instrument typically rely on a qualitative inspection technique to identify symbolic implications of drawing characteristics for aspects of personality functioning. As in the case of interpreting the DAP in the tradition of Machover, many of the interpretive hypotheses for the HTP suggested by Buck and by Hammer are quite specific. A door that is tiny in relation to the size of the house is interpreted as

indicating reluctance to make contact with the environment and an inhibited capacity for social relations, for example, and overemphasis on the roots of the tree where they make contact with the ground is taken as evidence of subjects' concerns about losing their grip on reality.

4.15.7.8.2 Psychometric foundations

The entire psychological literature contains only a handful of articles bearing on the psychometric foundations of the HTP. Most of these are over 25 years old, and none of them provides convincing supportive evidence for the interpretive uses of the instrument recommended by its leading proponents. Buck's (1985) 350-page revised manual contains extensive guidelines and case illustrations to facilitate interpretation, but neither reliability nor validity appears in the index. As intriguing as the rationale for the instrument may be to clinicians, especially those who are psychodynamically oriented, there is at present no empirical basis to warrant inferring personality characteristics from it.

4.15.7.8.3 Clinical utility

Similar to the DAP, the HTP offers the potential advantages in clinical practice of a brief, easily administered, nonverbal, and largely culture-free assessment instrument, along with perhaps being even less anxiety provoking than the DAP. Emerging refinements with the DAP suggest that the HTP as well might prove useful in identifying maladjustment in general, monitoring progress and change in psychotherapy, and even in pointing to possible specific areas of conflict and concern. Moreover, like the DAP inquiry, the HTP interrogation can produce story-telling content that in turn can be codified and suggest topics for further exploration. As matters presently stand, however, any such utility remains an unfulfilled potential, and examiners should be circumspect about including the HTP in their test batteries and basing any firm conclusions on it.

4.15.7.9 Kinetic Family Drawing

Machover (1948) and numerous other clinicians who pioneered in using the DAP to assess young people suggested that useful information might also be obtained by asking individuals to draw members of their family. This suggestion was formalized by Burns and Kaufman (1970, 1972) as the kinetic family drawing (KFD) technique, in which subjects are instructed to draw a picture of everyone in their family, including themselves, doing something. These

drawings are then examined for such objective features as omissions of body parts or of members of the family and interpreted according to the actions, styles, and symbols represented in them.

Actions in the Burns and Kaufman approach refer to the ways in which the figures drawn are behaving toward each other, which are thought to provide clues to the intensity and emotional tone of their relationships. Styles concern barriers between family members that prevent them from interacting at all, which may be expressed by drawing some of them at a far distance from the others or encasing them in a circle or a box. Symbols comprise a list of specific items, such as beds, flowers, stoves, cats, and the like, the inclusion of which is considered to reflect various specific unconscious impulses or concerns.

Publication of the KFD was followed by a school adaptation of the technique by Prout and Phillips (1974), known as the kinetic school drawing (KSD), in which children are asked to draw a school picture of themselves, their teacher, and a friend or two in which everyone is doing something. The KSD was intended to provide information about peer relationships and about attitudes and concerns related to school in the same manner as the KFD does for family relationships and feelings about the home.

Knoff and Prout (1985a, 1985b) subsequently recommended combining the KFD and KSD and administering both measures for purposes of analysis and comparison. This combined approach, which they call the kinetic drawing system, is expected to identify adjustment difficulties both at home and in school, to clarify causal or reciprocal relationships between family and school-related issues, and to indicate which people in subjects' lives (e.g., father, sister, teacher) are sources of support or tension. In a further extension of this approach, Burns (1987) has developed a kinetic house-tree-person, in which subjects are instructed to draw on a single page a picture of a house, a tree, and a person in "some kind of action."

4.15.7.9.1 Administration and scoring

Similar to procedures followed in other drawing techniques, administration of the KFD consists of giving subjects an 8.5 × 11 inch piece of plain paper and asking them to draw a picture of everyone in their family, including themselves, doing something. Beginning with an interpretive approach modeled basically after Machover's qualitative system, Burns (1982) developed a long list of actions, style, and symbol characteristics for examiners

to consider. As summarized by Knoff and Prout (1985a, 1985b), at least four objective methods for coding characteristics of kinetic drawings have also been proposed by various investigators, but none of these has become consistently visible in the literature. A review of the KFD literature by Handler and Habenicht (1994) indicates in general that cumulative knowledge concerning the adequacy and utility of this instrument has been limited by considerable variation in whether and how it has been scored in research studies and applied in clinical practice.

Also of note is a thematic elaboration of the KFD proposed by McConaughy and Achenbach (1994) and called the semistructured interview protocol. In this approach subjects are given the following question to answer after they have completed their family drawing: (i) what are they doing, (ii) what kind of person is [each member], (iii) what are three words that describe [each member], (iv) how does [each member] feel in this picture, (v) what is [each member] thinking, (vi) who do you get along with best, (vii) who do you get along with least, and (viii) what is going to happen next in your picture?

4.15.7.9.2 Psychometric foundations

Handler and Habenicht (1994) were able to reference a substantial number of publications concerning the KFD. However, they were forced to conclude that, despite its widespread use, this projective method has not yet been adequately developed with respect to its psychometric properties. Most of the systems previously proposed for coding KFD characteristics can achieve substantial inter-rater reliability, with percentages of agreement generally ranging well above 0.85 in various studies (Cummings, 1986; Handler & Habenicht, 1994). However, none of these scoring systems has demonstrated satisfactory retest reliability, even after very brief intervals, and there is little empirical basis for challenging the opinion that "the KFD still remains primarily a clinical instrument with inadequate norms and questionable validity" (Handler & Habenicht, 1994, p. 441).

With this in mind, Handler and Habenicht have recommended a holistic, integrative approach to KFD interpretation, much in the manner of the previously noted integrative DAP method used by Tharinger and Stark (1990), rather than the coding and summation of lists of signs and symbols. In the same study in which Tharinger and Stark demonstrated the superiority of their integrative DAP to the Koppitz scoring system, they also found that a holistic

method of evaluating kinetic drawings discriminated adjusted from maladjusted children more effectively than a traditional scoring guide for KFD interpretation developed by Reynolds (1978).

In developing his scoring guide, Reynolds urged clinicians and researchers to avoid cookbook approaches that fail to go beyond positing direct and absolute links between individual drawing characteristics and specific personality features. Instead, scores for drawing characteristics should be only a first step in interpreting drawings as gestalts that take on meaning only in relation to a youngster's personal, family, and cultural context. It can be hoped that eventual attention to this sound advice, perhaps through the standardization of integrated coding systems, will establish for kinetic drawings the validity that is presently lacking for them to be used with confidence in clinical evaluations.

4.15.7.9.3 Clinical utility

Conceived as a means of understanding how children and adolescents conceptualize their family and perceive themselves within the family context, the KFD is a potentially useful instrument for elaborating the interpersonal dynamics of young people and orchestrating individual or family therapy for those with adjustment difficulties. By employing the kinetic drawing system, examiners can explore school-based as well as family-related issues for such purposes. Of further potential benefit, subjects from a variety of ethnic and minority group backgrounds have been found to express dimensions of their family culture in their KFDs (Handler & Habenicht, 1994), which suggests a role for this instrument in multi-cultural assessment.

Unfortunately, however, the previously noted psychometric limitations of the KFD makes its clinical use problematic at present, particularly with respect to basing any firm conclusions on what and how subjects draw. Until such time as adequate research has clarified what the KFD and other projective drawings methods can and cannot do, clinicians are well advised to regard their figure drawing findings as suggesting but not confirming any notions about the subject. We would endorse in this regard the conclusion of Knoff (1990) that "The hypothesis generating use of projective drawings and their ability to be interpreted within various psychological orientations remains both viable and defensible [but] the validity of hypotheses tied to specific drawing characteristics must still be determined" (p. 99).

4.15.7.10 Sentence Completion Methods

Sentence completion methods consist of initial words or phrases, called stems, that subjects are asked to extend either orally or in writing into complete sentences. Typically used stems vary in length and structure from just one or two words, such as "People ..." or "I wish ...," to detailed specification of people or situations, such as "If only my mother would ..." or "When he found he had failed the examination, he ..." As in the case of responses on other projective methods, the manner in which subjects complete sentences is expected to provide an indirect source of information concerning their underlying feelings, attitudes, and level of adjustment. In addition to eliciting general information in these respects, the specificity in many sentence completion stems encourages subjects to reveal their orientation toward particular events and circumstances in their lives.

Sentence completion methods originated in word association tests dating back to the 1890s, and, as reviewed by Haak (1990) and Lah (1989b), were developed during the 1940s and 1950s into a large number of formal and informal versions, of which the most noteworthy published scales were the Rohde, the Sacks, the Forer, and the Miale-Holsopple sentence completion tests and the Rotter incomplete sentences blank (RISB). Of these, the RISB has become the best known and most widely used and includes an adult, a college, and a high school form (Rotter, Lah, & Rafferty, 1992). Also available are more recently developed sentence completion forms constructed by Brown and Unger (1992) for use with adolescents and adults and by Hart (1986) for evaluating school age children. Further comments focus primarily on the RISB, as the predominant and most recently revised exemplar of the sentence completion method, and on the Hart sentence completion test (HSCT), because of its specificity for assessing young people.

4.15.7.10.1 Administration and scoring

The RISB is a 40-item test printed on the front and back of a one-page form and comprised mainly of brief stems (eight have just a single word, 21 have two words, six have three words, and the remaining five have four words). Subjects are given a pencil and asked to "complete these sentences to express your real feelings." The test can be administered either individually or in a group setting; however, Rotter et al. (1992) caution against taking oral rather than written responses, primarily because

doing so injects an interpersonal component into the testing administration that can influence subjects' responses in various confounding ways.

The RISB and other sentence completion tests are typically interpreted in clinical practice by the inspection methods we have described previously; that is, examiners read the content of the items and from impressions of what they signify concerning a subject's probable personality characteristics. Rotter et al. (1992) also provide a scoring system for rating each item on a seven-point scale from 0 (most positive adjustment) to 6 (most indication of conflict). These ratings are totaled to yield an overall adjustment score. There is little indication that this or any other codification has received much attention in clinical practice or research studies, even though some work with college students indicated that the overall RISB score can discriminate those who are receiving counseling or psychotherapy from their nonpatient peers (Lah, 1989a).

The HSCT is also a 40-item measure in which the item stems were designed specifically to use with children and to sample family, social, school, and self dimensions of their lives (Hart, Kehle, & Davies, 1983). HSCT scoring involves rating each item as negative, neutral, or positive with respect to adjustment and also rating 10 scales composed of various clusters of items concerned with perceptions of self, family, and school on a five-point negative-to-positive continuum. Criteria are provided for each item to guide examiners in their ratings and thereby to enhance scoring objectivity and the prospects for achieving inter-rater agreement.

4.15.7.10.2 Psychometric foundations

Although there is a substantial RISB research literature, most of the published studies have used the instrument as a measure of adjustment but, as pointed out by Lah (1989a), they were not designed to evaluate its properties. Particularly with respect to its validity, there has been virtually no accumulation of empirical evidence to support any diagnostic or predictive inferences about personality functioning beyond the modestly demonstrated capacity of the RISB to identify maladjustment. Likewise, little systematic progress has been made in assessing the reliability of RISB data and establishing normative standards for them.

This is a regrettable state of affairs because, aside from being a psychodynamically compelling method of enriching personality assessment, the RISB is an eminently codeable instrument for which excellent interscorer agreement has been demonstrated using the

Rotter et al. approach (Lah, 1989a). Moreover, as discussed in the introductory portions of this chapter, sentence completion responses have numerous objective, structural, and behavioral components (e.g., time to completion, length of sentences, frequency of self- vs. other-reference, items omitted) that have rarely been considered in codifying these methods or attempting to establish a psychometric foundation for them.

The HSCT coding system has also achieved substantial inter-rater agreement and some preliminary success differentiating between emotionally disturbed and well-adjusted children. However, Hart (1986, p. 269) observed that it remained to further research to demonstrate adequate reliability and criterion validity for his instrument and to develop sufficiently broad normative data. The available literature, circa 1996, suggests that not much if any, progress has been made in this regard.

4.15.7.10.3 Clinical utility

Sentence completion methods bring to the personality assessment battery an easily administered projective test that can be given to groups as well as individuals and requires only about 30 minutes on average for subjects to complete. The scorable 40-item RISB high school form is suitable for high school and most middle school youngsters, and the 40-item HSCT can be used comfortably with younger children. Barring a language or reading difficulty, the meaning of the sentence completion stems is clear, and subjects are not asked to explain, elaborate, or otherwise account for their responses. Hence, incomplete sentence methods are often less threatening or anxiety-provoking than inkblot, story-telling, and figure drawing procedures. However, the previously noted relative lack of ambiguity in sentence completion stems means that subjects are more aware than on other projective methods of how they are presenting themselves through their responses.

At this point in time, the RISB and HSCT can be used with justification to form general impressions of a young person's level of adjustment and to formulate hypotheses concerning possible conflicts or concerns the young person has and how he or she may feel about self, other people, and certain situations. Such hypotheses can justifiably be expressed only as speculations, however, not as conclusions, and they should be considered reasonably correct only to the extent that they are supported by other reliable sources of information.

Haak (1990) has asserted that sentence completion methods can be used effectively

with children to rule out intellectual difficulties, attention deficit disorder, and stress and to rule in depression, anxiety, thought disturbance, and defensiveness, and provides detailed clinical guidelines for doing so. These suggestions seem sensible in each case, but they are entirely unverified by empirical data. Hence, like other impressions formed by experienced clinicians, Haak's diagnostic guidelines provide an agenda for confirmatory research, but they should not be elevated from speculations to the status of conclusions until that empirical confirmation becomes available.

4.15.8 FUTURE DIRECTIONS

Projective methods have been extensively used, taught, and studied for many years, and it seems likely that they will continue in the future to be regularly included in test batteries for assessing personality functioning in young people. Survey data cited in this chapter document that over the last 30 years psychologists have continued with much the same frequency to apply these instruments in practice and utilize them in research. During this same period the psychometric adequacy and clinical contributions of projective methods have been regularly and vigorously challenged, and this chapter indicates that the uses to which some of these instruments are put often go beyond available justification in empirical data.

How do we account for the persistent use by clinicians of many such as yet unvalidated methods? One could say that there just happen to be legions of uninformed or unethical clinicians in practice who do not hesitate to employ useless methods if doing so serves their purpose in some way. However, it seems doubtful that a vast segment of the profession deserves to be tarred with such a broad brush of evil. Rather, at least the majority of professionals who use projective methods must have good reason on the basis of their clinical and individual case experience to believe sincerely that these methods provide valid and useful information about personality functioning in ways that have not yet been translated into supporting research data. One could then say that such clinical confidence in projective methods is based solely on illusory correlation. However, it seems doubtful that a vast segment of professional personality assessors could be so thoroughly deluded.

These observations suggest that projective techniques have been and will continue to be used because they yield valuable information and generate fruitful hypotheses in clinical assessments. Hopefully, adequate research

methods will lead the way toward improved standardization and codification of projective methods, that will in turn enhance the psychometric foundations on which they rest. As Kuehnle (1996) has pointed out, inappropriate use of projective instruments for purposes for which they have not been validated, such as identifying children as having been traumatized or sexually abused, violates ethical standards and risks causing harm to young people and their families.

More than any of the other projective methods, the Rorschach has had the benefit of rigorous attention to research methodology, as demonstrated in publications by Exner (1995) and Weiner (1995a) and, as indicated earlier in this chapter, this method presently rests on a solid psychometric foundation. What the future of projective methods needs is similar methodological attention to documenting the reliability of other techniques and their validation for various purposes. To say that sophisticated research methods are not applicable to projective methods is as unwarranted as asserting that these methods are by nature invalid. To say that projective test responses cannot be examined scientifically without detracting from their idiographic richness sells projective methods short and prevents them from realizing their full potential.

Along with these needs for improved research and a narrowed gap between data and practice, the most important future direction for projective testing of young people lies in developing adequately representative normative data that will facilitate age-specific and multicultural assessment. There is no lack of projective techniques, but there is a decided lack of reliable information concerning how children and adolescents of different ages, from diverse backgrounds, and with various kinds of personality strengths and weaknesses, should be expected to respond to them. In addition to the nearly empty coffers of cross-sectional normative data of these kinds, the cupboard is virtually bare with respect to longitudinal data indicating how young people are likely to change over time and with maturation in how they respond to projective tests. Collecting and disseminating such data is the number one agenda item for the future development of projective assessment of children and adults.

4.15.9 SUMMARY

Projective tests are methods of personality assessment in which some degree of ambiguity in the test stimuli or instructions creates opportunities for subjects to structure their

responses in terms of their individual personality characteristics, and thereby provide information about the nature of these characteristics. Although projective methods are accordingly more ambiguous and less structured than so-called objective methods, the differences between these methods are relative rather than absolute. All projective tests contain objective as well as subjective features and elicit responses that are representative as well as symbolic of behavior, and they differ from each other in the extent to which they are ambiguous.

Because of their relatively unstructured nature, projective tests measure personality functioning in subtle and indirect ways and tap underlying psychological characteristics at a less conscious level than relatively structured measures. Projective test data consequently provide valuable information about how people are likely to think, feel, and act that is difficult to obtain from objective assessment procedures, and they are also less susceptible than objective test data to the influence of test-taking attitudes.

Projective methods can be used to good effect with children and adolescents as well as adults. The basic interpretive conclusions and hypothesis that attach to projective test variables apply regardless of the age of the subject, provided that examiners determine the implications of their data in the light of normative developmental expectations. Surveys of clinical and school settings indicate that the projective instruments most frequently administered in evaluating young people are the Rorschach inkblot method, the thematic apperception test (TAT), the children's apperception test, the Roberts apperception test for children (RATC), the tell-me-a-story (TEMAS) test, the draw-a-person, the house-tree-person, the kinetic family drawing, and alternate forms of the sentence completion test.

For each of these nine projective tests, this chapter reviews their composition, administration, scoring, psychometric foundations, and clinical utility. Despite widespread utilization of these nine tests in clinical practice to draw conclusions about the personality characteristics, level of adjustment, and treatment needs of young people, only the Rorschach presently rests on a solid empirical foundation. Properly collected and interpreted, Rorschach data provide numerous demonstrably reliable and valid indices that facilitate differential diagnosis and treatment planning for children and adolescents with adjustment difficulties.

The RATC and TEMAS have shown through the development of standardized administration and scoring procedures that story-telling methods have the potential to achieve psycho-

metric respectability. However, further research on the RATC and TEMAS is needed with regard to multicultural and adolescent norms, respectively. With regard to the TAT, recently emerging psychometrically sound schemes for coding specific personality characteristics reflected in thematic content, such as SCORS, appear to provide a basis for empirical decision making.

The other measures reviewed have in various ways also shown potential to be codified and refined on the basis of empirical data, and there is reason to be hopeful that advances in research methodology will eventually close a currently regrettable gap between what is known for sure about these projective methods and what is frequently assumed to be true about them in clinical practice. Until this gap is narrowed, and especially until such time as more extensive normative and multicultural data become available, most clinical inferences from projective data should be regarded as hypotheses to be confirmed rather than as facts on which to base conclusions and recommendations.

4.15.10 REFERENCES

- Abraham, P. P., Lepisto, B. L., Lewis, M. G., & Schultz, L. (1994). Changes in Rorschach variables of adolescents in residential treatment: An outcome study. *Journal of Personality Assessment*, 62, 505-514.
- Alvarado, N. (1994). Empirical validity of the Thematic Apperception Test. *Journal of Personality Assessment*, 63, 59-79.
- Ames, L. B., Metraux, R. W., Rodell, J. L., & Walker, R. N. (1974). *Child Rorschach responses* (Rev. ed.). New York: Brunner/Mazel.
- Ames, L. B., Metraux, R. W., & Walker, R. N. (1971). *Adolescent Rorschach responses* (Rev. ed.). New York: Brunner/Mazel.
- Archer, R. P., Imhof, E. A., Maruish, M., & Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. *Professional Psychology*, 22, 247-252.
- Archer, R. P., & Krishnamurthy, R. (1993a). A review of MMPI and Rorschach interrelationships in adult samples. *Journal of Personality Assessment*, 61, 277-293.
- Archer, R. P., & Krishnamurthy, R. (1993b). Combining the Rorschach and MMPI in the assessment of adolescents. *Journal of Personality Assessment*, 60, 132-140.
- Atkinson, J. W., & Feather, N. T. (1966). *A theory of achievement motivation*. New York: Wiley.
- Bardos, A. N. (1993). Human figure drawings: Abusing the abused. *School Psychology Quarterly*, 8, 177-181.
- Bellak, L. (1993). *The T.A.T., C.A.T., and S.A.T. in clinical use* (5th ed.). Boston: Allyn & Bacon.
- Bellak, L., & Hurvich, M. (1966). A human modification of the Children's Apperception Test. *Journal of Projective Techniques*, 30, 228-242.
- Bellak, L., & Siegel, H. (1989). The Children's Apperception Test (CAT). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (Vol. II, pp. 99-127). Boston: Allyn & Bacon.
- Bornstein, R. F. (1995). Sex differences in objective and projective dependency tests: A meta-analytic review. *Assessment*, 2, 319-331.

- Buck, J. N. (1948). The H-T-P technique, a qualitative and quantitative method. *Journal of Clinical Psychology*, 4, 317-396.
- Buck, J. N. (1985). *The House-Tree-Person technique: Revised manual*. Los Angeles: Western Psychological Services.
- Burns, R. C. (1982). *Self-growth in families: Kinetic Family Drawings (K-F-D) research and application*. New York: Brunner/Mazel.
- Burns, R. C. (1987). *Kinetic-House-Tree-Person drawings (K-H-T-P)*. New York: Brunner/Mazel.
- Burns, R. C., & Kaufman, S. H. (1970). *Kinetic Family Drawings (K-F-D): An introduction to understanding children through kinetic drawings*. New York: Brunner/Mazel.
- Burns, R. C., & Kaufman, S. H. (1972). *Actions, styles, and symbols in Kinetic Family Drawings (K-F-D)*. New York: Brunner/Mazel.
- Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology*, 47, 87-111.
- Chandler, L. A. (1990). The projective hypothesis and the development of projective techniques for children. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (Vol. 2, pp. 55-69). New York: Guilford Press.
- Costantino, G., Colon-Malgady, G., Malgady, R. G., & Perez, A. (1991). Assessment of attention deficit disorder using a thematic apperception technique. *Journal of Personality Assessment*, 57, 97-95.
- Costantino, G., Malgady, R. G., Bailey, J., & Colon-Malgady, G. (1989). *Clinical utility of TEMAS: A projective test for children*. Paper presented at the meeting of the Society for Personality Assessment, New York.
- Costantino, G., Malgady, R. G., Colon-Malgady, G., & Bailey, J. (1992). Clinical utility of the TEMAS with non minority children. *Journal of Personality Assessment*, 59, 433-438.
- Costantino, G., Malgady, R. G., & Rogler, L. H. (1988). *TEMAS (Tell-Me-A-Story) manual*. Los Angeles: Western Psychological Services.
- Costantino, G., Malgady, R. G., Rogler, L. H., & Tusi, E. C. (1988). Discriminant analysis of clinical outpatients and public school children by TEMAS: A thematic apperception test for Hispanics and Blacks. *Journal of Personality Assessment*, 52, 670-678.
- Cramer, P. (1987). The development of defense mechanisms. *Journal of Personality*, 55, 597-614.
- Cramer, P., & Blatt, S. J. (1990). Use of the TAT to measure change in defense mechanisms following intensive psychotherapy. *Journal of Personality Assessment*, 54, 236-251.
- Cummings, J. A. (1986). Projective drawings. In H. M. Knoff (Ed.), *The assessment of child and adolescent personality* (pp. 199-204). New York: Guilford Press.
- Dana, R. H. (1985). Thematic Apperception Test (TAT). In C. Newmark (Ed.), *Major psychological assessment instruments* (pp. 89-134). Boston: Allyn & Bacon.
- Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn & Bacon.
- Dana, R. H. (1996). Culturally competent assessment practice in the United States. *Journal of Personality Assessment*, 66, 472-487.
- Di Leo, J. H. (1983). *Interpreting children's drawings*. New York: Brunner/Mazel.
- Elbert, J. C., & Holden, E. W. (1987). Child diagnostic assessment: Current training practices in clinical psychology internships. *Professional Psychology*, 18, 587-596.
- Exner, J. E., Jr. (1991). *The Rorschach: A comprehensive system. Vol. 2. Interpretation* (2nd ed.). New York: Wiley.
- Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system. Vol. 1. Basic foundations* (3rd ed.). New York: Wiley.
- Exner, J. E., Jr. (Ed.) (1995). *Issues and methods in Rorschach research*. Mahwah, NJ: Erlbaum.
- Exner, J. E., Jr., & Andronikoff-Sanglade, A. (1992). Rorschach changes following brief and short-term therapy. *Journal of Personality Assessment*, 59, 59-71.
- Exner, J. E., Jr., Thomas, E. A., & Mason, B. (1985). Children's Rorschachs: Description and prediction. *Journal of Personality Assessment*, 49, 13-20.
- Exner, J. E., Jr., & Weiner, I. B. (1995). *The Rorschach: A comprehensive system. Vol. 3. Assessment of children and adolescents* (2nd ed.). New York: Wiley.
- Finn, S. E. (1996, March). *Assessment feedback integrating MMPI-2 and Rorschach findings*. Paper presented at the annual meeting of the Society for Personality Assessment, Denver, CO.
- Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology*, 8, 389-413.
- Freeddenfeld, R., Ornduff, S., & Kelsey, R. M. (1995). Object relations and physical abuse: A TAT analysis. *Journal of Personality Assessment*, 64, 552-568.
- Freud, S. (1958). Psycho-analytic notes upon an autobiographical account of a case of paranoia (dementia paranoides). *Standard edition* (Vol. XII, pp. 9-82). London: Hogarth. (Original work published in 1911.)
- Freud, S. (1962). Further remarks on the neuro-psychoses of defence. *Standard edition* (Vol. III, pp. 162-185). London: Hogarth. (Original work published in 1896.)
- Ganellen, R. J. (1996). *Integrating the Rorschach and the MMPI-2 in personality assessment*. Mahwah, NJ: Erlbaum.
- Gardner, R. A. (1971). *Therapeutic communication with children: The mutual storytelling technique*. Northvale, NJ: Aronson.
- Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. New York: Harcourt, Brace & World.
- Gresham, F. M. (1993). "What's wrong in this picture?": Response to Motta et al.'s review of human figure drawings. *School Psychology Quarterly*, 8, 182-186.
- Haak, R. A. (1990). Using the sentence completion to assess emotional disturbance. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (Vol. 2, pp. 147-167). New York: Guilford Press.
- Hammer, E. F. (1958). *The clinical application of projective drawings*. Springfield, IL: Charles C. Thomas.
- Hammer, E. F. (1985). The House-Tree-Person test. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (pp. 135-164). Boston: Allyn & Bacon.
- Handler, L. (1985). The clinical use of the Draw-a-person test (DAP). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (pp. 165-216). Boston: Allyn & Bacon.
- Handler, L., & Habenicht, D. (1994). The Kinetic Family Drawing technique: A review of the literature. *Journal of Personality Assessment*, 62, 440-464.
- Harris, D. B. (1963). *Children's drawings as a measure of intellectual maturity*. New York: Harcourt, Brace & World.
- Hart, D. H. (1986). The sentence completion techniques. In H. M. Knoff (Ed.), *The assessment of child and adolescent personality* (pp. 245-272). New York: Guilford Press.
- Hart, D. H., Kehle, T. J., & Davies, M. V. (1983). Effectiveness of sentence completion techniques: A review of the Hart Sentence Completion Test. *School Psychology Review*, 12, 428-434.
- Haworth, M. (1965). *A schedule of adaptive mechanisms in CAT responses*. Larchmont, NY: CPS.
- Hibbard, S., Farmer, L., Wells, C., Difillipo, E., Barry, W., Korman, R., & Sloan, P. (1994). Validation of Cramer's defense mechanism manual for the TAT. *Journal of Personality Assessment*, 63, 197-210.

- Hibbard, S., Hilsenroth, M. J., Hibbard, J. K., & Nash, M. R. (1995). A validity study of two projective object representation measures. *Psychological Assessment*, 7, 432-439.
- Hoffman, S., & Kupperman, N. (1990). Indirect treatment of traumatic psychological experiences: The use of TAT cards. *American Journal of Psychotherapy*, 44, 107-115.
- Holtzman, W. H. (1993). An unjustified, sweeping indictment by Motta et al. of human figure drawings for assessing psychological functioning. *School Psychology Quarterly*, 8, 189-190.
- Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review*, 21, 271-284.
- Kahill, S. (1984). Human figure drawings in adults: An update of the empirical literature. *Canadian Psychology*, 25, 269-292.
- Kennedy, M. L., Faust, D., Willis, W. G., & Piotrowski, C. (1994). Social-emotional assessment practices in school psychology. *Journal of Psychoeducational Assessment*, 12, 228-240.
- Knoff, H. M. (1990). Evaluation of projective drawings. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (Vol. 2, pp. 89-146). New York: Guilford Press.
- Knoff, H. M., & Prout, H. T. (1985a). The Kinetic Drawing System: A review and integration of the Kinetic Family and Kinetic School Drawing techniques. *Psychology in the Schools*, 22, 50-59.
- Knoff, H. M., & Prout, H. T. (1985b). *The Kinetic drawing system: Family and school*. Los Angeles: Western Psychological Services.
- Kohlberg, L. (1976). Moral stage and moralization: The cognitive developmental approach. In T. Lickona (Ed.), *Moral development and behavior: Theory, research, and social issues* (pp. 31-53). New York: Holt, Rinehart & Winston.
- Koppitz, E. M. (1968). *Psychological evaluation of children's human figure drawings*. New York: Grune & Stratton.
- Koppitz, E. M. (1984). *Psychological evaluation of human figure drawings by middle school pupils*. New York: Grune & Stratton.
- Kuehnle, K. (1996). *Assessing allegations of child sexual abuse*. Sarasota, FL: Professional Resource Press.
- Lah, M. I. (1989a). New validity, normative, and scoring data for the Rotter Incomplete Sentences Blank. *Journal of Personality Assessment*, 53, 607-620.
- Lah, M. I. (1989b). Sentence completion tests. In C. S. Newmark (Ed.), *Major psychological assessment instruments, Volume II* (pp. 133-163). Boston: Allyn & Bacon.
- Leichtman, M. (1996). *The Rorschach: A developmental perspective*. Hillsdale, NJ: Analytic Press.
- Machover, K. (1948). *Personality projection in the drawing of the human figure*. Springfield, IL: Charles C. Thomas.
- Machover, K. (1951). Drawing of the human figure: A method of personality investigation. In H. H. Anderson & C. L. Anderson (Eds.), *An introduction to projective techniques* (pp. 341-369). New York: Prentice-Hall.
- McArthur, D. S., & Roberts, G. E. (1990). *Roberts Apperception Test for Children manual*. Los Angeles: Western Psychological Services.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. New York: Appleton-Century-Crofts.
- McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, 96, 690-702.
- McConaughy, S. H., & Achenbach, T. M. (1994). *Manual for the semistructured clinical interview for children and adolescents*. Burlington, VT: University of Vermont Department of Psychiatry.
- McDowell, C., & Acklin, M. W. (1996). Standardizing procedures for calculating Rorschach inter-rater reliability: Conceptual and empirical foundations. *Journal of Personality Assessment*, 66, 308-320.
- McNeish, T. J., & Naglieri, J. A. (1991). Identification of the seriously emotionally disturbed using the Draw A Person: Screening Procedure for Emotional Disturbance. *Journal of Special Education*, 27, 115-121.
- Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry*, 34, 289-306.
- Motta, R. W., Little, S. G., & Tobin, M. I. (1993). The use and abuse of human figure drawings. *School Psychology Quarterly*, 8, 162-169.
- Murray, H. A. (1943). *Thematic Apperception Test manual*. Cambridge, MA: Harvard University Press.
- Murray, H. A. (1951). Uses of the Thematic Apperception Test. *American Journal of Psychiatry*, 107, 577-581.
- Murray, H. A. (1971). *Thematic Apperception Test: Manual* (Rev. ed.). Cambridge, MA: Harvard University Press.
- Murstein, B. I. (1963). *Theory and research in projective techniques (Emphasizing the TAT)*. New York: Wiley.
- Murstein, B. I., & Wolf, S. R. (1970). Empirical test of the "levels" hypothesis with five projective techniques. *Journal of Abnormal Psychology*, 75, 38-44.
- Naglieri, J. A. (1988). *Draw A Person: A quantitative scoring system*. New York: Psychological Corporation.
- Naglieri, J. A., McNeish, T. J., & Bardos, A. N. (1991). *Draw-A-Person: Screening Procedure for Emotional Disturbance*. Austin, TX: ProEd.
- Naglieri, J. A., & Pfeiffer, S. I. (1992). Performance of disruptive behavior disordered and normal samples on the Draw A Person: Screening Procedure for Emotional Disturbance. *Psychological Assessment*, 4, 156-159.
- Obrzut, J. E., & Boliek, C. A. (1986). Thematic approaches to personality assessment with children and adolescents. In H. M. Knoff (Ed.), *The assessment of child and adolescent personality* (pp. 173-198). New York: Guilford Press.
- Ornduff, S. R., Freedenfeld, R., Kelsey, R. M., & Critelli, J. (1994). Object relations of sexually abused female subjects: A TAT analysis. *Journal of Personality Assessment*, 63, 223-228.
- Ornduff, S. R., & Kelsey, R. M. (1996). Object relations of sexually and physically abused female children: A TAT analysis. *Journal of Personality Assessment*, 66, 91-105.
- Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin*, 103, 367-373.
- Piotrowski, C. (1996). The status of Exner's Comprehensive System in contemporary research. *Perceptual and Motor Skills*, 82, 1341-1342.
- Piotrowski, C., & Keller, J. W. (1989). Psychological testing in outpatient facilities: A national study. *Professional Psychology*, 20, 423-425.
- Prout, H. T., & Phillips, P. D. (1974). A clinical note: The kinetic school drawing. *Psychology in the Schools*, 11, 303-306.
- Reynolds, C. R. (1978). A quick scoring guide to the interpretation of children's Kinetic Family Drawings (KFD). *Psychology in the School*, 15, 489-492.
- Roback, H. B. (1968). Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. *Psychological Bulletin*, 70, 1-19.
- Ronan, G. F., Colavito, V. A., & Hammontree, S. R. (1993). Personal problem-solving system for scoring TAT responses: Preliminary validity and reliability data. *Journal of Personality Assessment*, 61, 28-40.
- Rorschach, H. (1942). *Psychodiagnostics*. Bern, Switzerland: Hans Huber. (Original work published in 1921.)
- Rotter, J. B., Lah, M. I., & Rafferty, J. E. (1992). *Manual—Rotter Incomplete Sentences Blank* (2nd ed.). Orlando, FL: Psychological Corporation.

- Seaton, B., & Allen, J. (1996, March). *Interscorer reliability of Rorschach structural summary data*. Poster session presented at the annual meeting of the Society for Personality Assessment, Denver, CO.
- Smith, D., & Dumont, F. (1995). A cautionary study: Unwarranted interpretations of the Draw-A-Person test. *Professional Psychology, 3*, 298–303.
- Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12*, 331–350.
- Stone, H. K., & Dellis, N. P. (1960). An exploratory investigation into the levels hypothesis. *Journal of Projective Techniques, 24*, 333–340.
- Sutton, P. M., & Swenson, C. H. (1983). The reliability and concurrent validity of alternative methods for assessing ego development. *Journal of Personality Development, 47*, 468–475.
- Swensen, C. H. (1957). Empirical evaluations of human figure drawings. *Psychological Bulletin, 54*, 431–466.
- Swensen, C. H. (1968). Empirical evaluations of human figure drawings: 1957–1966. *Psychological Bulletin, 70*, 20–44.
- Tharinger, D. J., & Stark, K. (1990). A qualitative versus quantitative approach to the Draw-a-Person and Kinetic Family Drawing: A study of mood- and anxiety-disorder children. *Psychological Assessment, 2*, 365–375.
- Thomas, A. D., & Dudek, S. Z. (1985). Interpersonal affect in TAT responses: A scoring system. *Journal of Personality Assessment, 49*, 30–37.
- Vane, J. R. (1981). The Thematic Apperception Test: A review. *School Psychology Review, 1*, 319–336.
- Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology, 26*, 54–60.
- Weiner, I. B. (1977). Projective tests in differential diagnosis. In B. B. Wolman (Ed.), *International encyclopedia of neurology, psychiatry, psychoanalysis, and psychology* (pp. 112–116). Princeton, NJ: Van Nostrand Reinhold.
- Weiner, I. B. (1986). Assessing children and adolescents with the Rorschach. In H. M. Knoff (Ed.), *The assessment of child and adolescent personality* (pp. 141–171). New York: Guilford Press.
- Weiner, I. B. (1993). Clinical considerations in the conjoint use of the Rorschach and the MMPI. *Journal of Personality Assessment, 60*, 148–152.
- Weiner, I. B. (1994). Rorschach assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome evaluation* (pp. 249–278). Hillsdale, NJ: Erlbaum.
- Weiner, I. B. (1995a). Methodological considerations in Rorschach research. *Psychological Assessment, 7*, 330–337.
- Weiner, I. B. (1995b). Psychometric issues in forensic applications of the MMPI-2. In Y. S. Ben-Porath, J. R. Graham, G. C. N. Hall, R. D. Hirschman, & M. S. Zaragoza (Eds.), *Forensic applications of the MMPI-2* (pp. 48–81). Thousand Oaks, CA: Sage.
- Weiner, I. B. (1996). Some observations on the validity of the Rorschach Inkblot Method. *Psychological Assessment, 8*, 206–213.
- Weiner, I. B. (1997). Current status of the Rorschach Inkblot Method. *Journal of Personality Assessment, 68*, 5–19.
- Weiner, I. B., & Exner, J. E., Jr. (1991). Rorschach changes in long-term and short-term psychotherapy. *Journal of Personality Assessment, 56*, 453–465.
- Westen, D., Klepser, J., Ruffins, S. A., Silverman, M., Lifton, N., & Boekamp, J. (1991). Object relations in childhood and adolescence: The development of working representations. *Journal of Consulting and Clinical Psychology, 59*, 400–409.
- Westen, D., Lohr, N., Silk, K., Kerber, K., & Goodrich, S. (1985). *Object relations and social cognition TAT scoring manual*. Ann Arbor, MI: University of Michigan.
- Westen, D., Ludolph, P., Block, M. J., Wixom, J., & Wiss, F. C. (1990). Developmental history and object relations in psychiatrically disturbed adolescent girls. *American Journal of Psychiatry, 147*, 1061–1068.
- Westen, D., Ludolph, P., Lerner, H., Ruffins, S., & Wiss, F. C. (1990). Object relations in borderline adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 29*, 338–348.
- Worchel, F. F., Rae, W. A., Olson, T. K., & Crowley, S. L. (1992). Selective responsiveness of chronically ill children to assessments of depression. *Journal of Personality Assessment, 59*, 605–615.

4.16

Assessment of Schema and Problem-solving Strategies with Projective Techniques

HEDWIG TEGLASI

University of Maryland, College Park, MD, USA

4.16.1 INTRODUCTION	460
4.16.2 PATTERNS OF USE	460
4.16.3 BASIC ASSUMPTIONS OF PROJECTIVE TECHNIQUES AND RESEARCH FROM OTHER PSYCHOLOGY SUBFIELDS	461
4.16.3.1 <i>Schema Theory and Projective Techniques</i>	462
4.16.3.2 <i>Unconscious Processing of Information</i>	464
4.16.3.3 <i>Inert Knowledge vs. Usable Knowledge</i>	465
4.16.3.4 <i>Affective and Motivational Influences on Cognition</i>	465
4.16.3.5 <i>Narrative Psychology and Thematic Apperception Techniques</i>	466
4.16.4 LEVELS OF PERSONALITY ASSESSED BY PROJECTIVE AND QUESTIONNAIRE METHODS	467
4.16.5 SELF-REPORT VS. PROJECTIVE PERSONALITY TESTING	470
4.16.6 CONTRIBUTION OF PROJECTIVE TECHNIQUES TO DIAGNOSIS AND INTERVENTION	470
4.16.7 PROJECTIVE TECHNIQUES AS PERFORMANCE MEASURES OF PERSONALITY	472
4.16.7.1 <i>Total Personality</i>	474
4.16.7.2 <i>Many Correct Solutions</i>	476
4.16.7.3 <i>Generalize to Different Criteria</i>	476
4.16.7.4 <i>Differences in Conditions of Learning and Performance</i>	477
4.16.8 SPECIFIC PROJECTIVE TECHNIQUES	477
4.16.8.1 <i>Stimulus</i>	477
4.16.8.2 <i>Response</i>	477
4.16.8.3 <i>Task Demand</i>	478
4.16.8.4 <i>Interpretation</i>	478
4.16.9 THEMATIC APPERCEPTIVE TECHNIQUES	479
4.16.9.1 <i>Response</i>	479
4.16.9.2 <i>Stimuli</i>	479
4.16.9.3 <i>Interpretation</i>	480
4.16.10 RORSCHACH TECHNIQUE	481
4.16.10.1 <i>Stimulus Features and Response Parameters</i>	482
4.16.10.2 <i>Interpretation</i>	483
4.16.11 DRAWING TECHNIQUES	483
4.16.11.1 <i>Interpretation</i>	484
4.16.12 CASE ILLUSTRATION	484

4.16.13	VALIDATION ISSUES	487
4.16.13.1	Reliability	488
4.16.13.1.1	Scorer reliability	488
4.16.13.1.2	Decisional reliability	488
4.16.13.1.3	Test-retest reliability	488
4.16.13.1.4	Internal consistency	488
4.16.13.2	Construct Validation	489
4.16.13.3	Multitrait, Multimethod Validation	489
4.16.13.4	Criterion Validation	489
4.16.13.5	Part-Whole Relationships	490
4.16.13.6	Convergence of Psychometric and Conceptual Treatment of Data	490
4.16.13.6.1	Normative	491
4.16.13.6.2	Nomothetic	491
4.16.13.6.3	Case study	491
4.16.14	FUTURE DIRECTIONS	492
4.16.15	SUMMARY	493
4.16.16	REFERENCES	495

4.16.1 INTRODUCTION

Clinical assessment practices typically lag behind the conceptualizations psychologists use to understand adaptive and maladaptive patterns in human functioning. We recognize that personality is a functional whole, but cling to assessment procedures that resemble the proverbial blind men and the elephant. Each method of assessment pertains to one dimension or level of personality, yet we expect different sources of data to confirm one another rather than to reveal the different patterns contributing to the larger mosaic. A clarification of the part-whole relationships with regard to personality assessment promotes an understanding of the place of projective techniques in a comprehensive evaluation. The aims of this chapter are: (i) to facilitate the process of integration of projective techniques with self-report approaches to personality assessment; (ii) to reframe projective techniques as performance measures of personality that differ in their demands from other tasks within the battery; (iii) to show that the central tenets of projective techniques embodied in the projective hypothesis are supported by empirical research in various psychology subfields; and (iv) to demonstrate the utility of projective techniques for assessing schema (the internal representation of reality) and problem-solving strategies to meet the reality demands posed by the task.

That projective techniques do not belong to a particular theory of personality was asserted early on (Auld, 1954). Projective methods yield open-ended responses, and their interpretation can accommodate changing paradigms in various subdisciplines, such as the cognitive and neurosciences as well as in the study of emotion and social cognition. It is easy to say that the framework to understand and interpret responses to projective tasks should coincide with

key concepts in the various subdisciplines, but the task of continuously updating the conceptual frameworks that are applied to projective methods remains an ongoing challenge.

Responses to projective tests yield information about the internal structure of personality and provide the context for understanding overt behavior and symptomatology. This emphasis on inner structure is currently shared by social, cognitive, and clinical psychology. Newer appreciation is emerging of the wide role of emotions and unconscious processes (e.g., Bargh, 1994; Clore, Schwarz, & Conway, 1994), as well as of meaning structures (e.g., Bruner, 1986) underlying behavioral expression. Projective methods reveal the structure and process of personality that are largely unconscious, and they differ in important ways from checklists or self-report interviews. In addition, the competencies required by projective tasks can be distinguished from those demanded by other measures in the typical assessment battery such as the Wechsler scales.

Therefore, projective techniques can be viewed as performance measures of personality that reveal inner structures or schema as resources needed in daily life. Each projective method sets a task that requires the application of these structures to respond adaptively. All factors contributing to personality development are involved in shaping these inner structures of meaning. Therefore, the interpretation of responses to projective techniques promotes the integration of knowledge from various subdisciplines.

4.16.2 PATTERNS OF USE

Despite criticisms of projective tests, there continues to be broad interest in the Rorschach and the Thematic Apperception Test (TAT;

Butcher & Rouse, 1996). Most clinical psychology doctoral training programs include formal instruction in the Rorschach (Piotrowski & Zalewski, 1993). In addition, the majority of internship sites approved by the American Psychological Association (APA) value knowledge in Rorschach technique (Durand, Blanchard, & Mindell 1988). The Exner Comprehensive System (Exner, 1993; Exner & Weiner, 1995) has become widely accepted because it provides a more reliable and objective basis for interpretation than was previously available. Surveys find the Rorschach, TAT, and various drawing methods among the top 10 most frequently used assessment techniques (Archer, Marnish, Imhof, & Piotrowski, 1991; Piotrowski & Keller, 1989; Piotrowski, Sherry, & Keller, 1985; Watkins, Campbell, & McGregor, 1988; Watkins, Campbell, Nieberding, & Hallmark, 1995). Interest in the Rorschach and TAT is also evident outside of the USA (Weiner, 1994).

The contrast between the negative evaluation of projective methods by researchers and their continued popularity with practitioners (Piotrowski, 1984) begs for a scrutiny of the assumptions and methods applied to judging the utility of these techniques. This state of affairs calls for an analysis not only of psychometric properties but also of the conceptual and methodological underpinnings pertaining to projective testing. Experienced practitioners continue to use these techniques because they find them helpful. For instance, an advantage of the Rorschach is that the validity of its administration is independent of the client's reading level (Archer et al., 1991). Shneidman's (1951) edited volume presented the conclusions of 16 clinicians who blindly interpreted the same TAT protocol using their own preferred method. Despite different approaches and foci, they were remarkably accurate and consistent with each other. Also included in this volume were Klopfer's conclusions based on a blind interpretation of the same individual's Rorschach record which was highly consistent with those based on the TAT protocols. It should be noted that the interpreters all had one thing in common—they were "experts." In light of what we know about how experts perform (e.g., Chi, Glaser, & Farr, 1987), we would not expect to replicate such findings with more recently trained individuals.

In evaluating the evidence of reliability and validity in the clinical use of the TAT, the training and experience of interpreters was a primary consideration (Karon, 1981). Research designs that isolate response variables and subject them to separate statistical analyses do not mirror the interpretation of projective

methods in clinical practice. The clinician simultaneously evaluates multiple dimensions of the response in reference to a psychological construct rather than adding individual response variables. More sophisticated validation efforts are needed that focus on the integrative approach rather than piecemeal summation of isolated response characteristics (e.g., Handler & Habenicht, 1994). Finally, more refined approaches to validation suggest that different facets of psychological constructs are measured by different techniques. For instance, a comprehensive meta-analysis indicated that projective measures of achievement motivation predict long term behavioral trends, whereas self-report measures predict immediate choices (Spangler, 1992). Refinements in conceptualization and approaches to validation are calling into question earlier criticism based on psychometric shortcomings.

4.16.3 BASIC ASSUMPTIONS OF PROJECTIVE TECHNIQUES AND RESEARCH FROM OTHER PSYCHOLOGY SUBFIELDS

The term "projective technique" was coined by Frank (1939) to describe tests using ambiguous stimuli and/or tasks that are less obvious in their intent and are therefore less subject to faking. The aim of any projective technique is to set a task that allows the individual to express characteristic ways of perceiving and organizing experiences. The uniqueness of the individual's style of responding is particularly evident when stimuli are ambiguous and there is no ready response. Under such conditions, there are many possible ways to approach the task or situation, and the person must actively organize the response. A central assumption of all projective methods is that stimuli from the environment are perceived and organized by the individual's specific needs, motives, feelings, perceptual sets, and cognitive structures, and that in large part this process occurs automatically and outside of awareness (Frank, 1948).

These assumptions about projective testing, first elaborated by Frank (1939) in his analysis of the projective hypothesis, are compatible with converging trends across various subfields in psychology pointing to the role of previously organized inner structures or mental sets in the interpretation of stimuli. Since the cognitive revolution of the 1960s, psychology has undergone a significant transformation by moving away from an emphasis on stimulus-response units to describe human perception and action to increasing emphasis on how humans impose

meaning and organization on their experiences (Singer & Salovey, 1991). Within cognitive psychology, simple associational models (e.g., paired associate learning) have become less important in describing ongoing information processing than the organized meaning structures, such as schema or scripts that are brought to bear on the interpretation of experiences (Piaget, 1954; Taylor & Crocker, 1981). At the same time, developments in social psychology (Abelson, 1976; Festinger, 1957; Heider, 1958; Kelley, 1967) led to the exploration of organized meaning structures as the fundamental bases of social action (Wyer & Srull, 1994). In the psychodynamic paradigm, the emergence of object relations theory (Blatt & Wild, 1976) implicitly relies on internal structures similar to schema to guide information processing about the self and other. The considerable work in cognitive, social, personality, and clinical psychology bearing on schema theory and supporting the projective hypothesis is briefly highlighted next.

4.16.3.1 Schema Theory and Projective Techniques

Earlier cognitive work within clinical psychology focused on the role of self-statements and inner thoughts. But clinical scientists interested in the role of cognition have increasingly adopted an information-processing approach to clinical phenomena (Ingram, 1986). This approach focuses on the schema that provide rules guiding behavior in social relationships and appear to influence how information about relationships is stored in memory (Fiske, Haslam, & Fiske, 1991). Well-learned, organized knowledge structures (schema or scripts) have considerable impact on memory storage and retrieval. Projective tasks require the superimposition of previously acquired schematic structures to the perception of stimuli and the organization of the response. Acklin (1994) presented a reformulation of the responses to the Rorschach on the basis of schema theory and information processing. Within this framework, emphasis is placed on the nature and adaptiveness of the schema activated by the stimulus properties.

Memory structures such as schema or scripts that store information about situations, persons, and events guide the interpretation of experiences by providing criteria for regulating attention to lend focus to the process of encoding, storage, and retrieval of information in specific domains (Taylor & Crocker, 1981). Event schema refer to the type of script that organizes understanding of a sequence of events

in a stereotyped or routine situation such as ordering dinner in a restaurant (Abelson, 1981). Such a "script is a set of expectations about what will happen next in a well understood situation" (Schank, 1990, p. 7). As long as we know what script others are following, we know how to act and to predict the actions of others. Upon entering a restaurant, we are seated, order from the menu, receive our meal, pay, and depart. This process of comparing experience with the internal schema enables us to perceive rapidly unstated information and to anticipate what will happen next. Some individuals may view situations for which they have no scripts (novel) as challenging, whereas others are stymied (e.g., "speechless") in such conditions.

Person schema include information about the self, others, and their interactions. They deal with the individual's rules for predicting, interpreting, responding to, and controlling affectively charged encounters (Tomkins, 1979). For Tomkins (1979, 1987, 1991), emotion is a key organizer of personal scripts. Once the script has been formed, the script organizes and modifies new experiences to fit the preexisting structure. Person schema develop through the individual's synthesis of past experiences according to the individual's style of processing information (Horowitz, 1991). Therefore the construct of person schema brings together models of perception, cognition, memory, affect, action, and feedback. Furthermore, it has the potential to integrate psychodynamic formulations and cognitive perspectives in understanding personality (Horowitz, 1991; Stein & Young, 1992). This process of interpreting new information according to previously acquired sets takes place without conscious awareness and may also lead to systematic distortions in perception, interpretation, and action in interpersonal encounters.

Causal schema give rise to attributions about the causes of events. Schema theory and attribution theory overlap in that attribution theory maintains that individuals are motivated to make causal inferences about experiences and that attributions are made in ways that are congruent with existing schema of self and other, and of the assumed relationships between causes and effects (Tunis, 1991). The relationship between such causal attributions and depression has been studied.

Evidence for the existence of schema is seen in their consequences for information processing, particularly in how information is remembered and retrieved. Schema-driven information processing involves the application of organized knowledge structures akin to theories that govern the perception and interpretation of facts. Thus, schema function to allow general

knowledge to influence perception of specific experiences. Schema increase efficiency in identifying perceptions, organizing them into manageable units, filling in missing information, and selecting a strategy for how to obtain further information if needed. Because schema guide behavior in adaptive and maladaptive ways, their study pertains to clinical psychology as well as to the general study of personality, cognition, and social perception. Understanding the recurrence of maladaptive interpersonal patterns is facilitated by studying the operation of schema along with the realistic properties of the current situation (Horowitz, 1991).

The overlap between schema or script theory and the projective hypothesis is evident in their common features. The development of schema or scripts and their retrieval from memory represent an unconscious process through which past perceptions influence the interpretation of current situations. These are precisely the processes that are assessed with projective techniques. The focus of projective methods is on the application of schema to the task demands (e.g., how schema influence the response process) and, by extension, to life challenges requiring similar adaptations.

Even though schema are structures of the mind, they must be sufficiently malleable to adapt to new situations and new configurations of events (Rummelhart, Smolensbus, McClelland, & Hinton, 1986). Therefore, it is appropriate to view schema as emerging when they are needed "from the interaction of large numbers of much simpler elements all working in concert with one another" (p. 20). A distinction has been made between enduring schema and working models (Horowitz, 1991). The former are intrapsychic meaning structures containing generalized formats of knowledge that can be activated by other mental activities related to that knowledge such as motivational concerns (Fiske & Taylor, 1984). Working models combine internal and external sources of information, such as when the individual is actively contemplating an interpersonal situation or task such as the TAT. Working models actively integrate stimuli from a current situation with past knowledge through triggering of an associative network of ideas drawn from enduring person schema.

A person's repertoire may include several different enduring person schema in relation to a given type of relationship, situation, or activity. The working model may incorporate different elements from each of the person schema. The wider the repertoire of enduring person schema, the more flexibility in constructing working models of interpersonal situations. Discrepancies between working

models and active, enduring schema may contribute to upsetting emotional experiences. Working models may not match the actual qualities of current social situations, leading to errors in judgment and behavior and to subsequent negative emotions. The aim of many psychotherapeutic techniques is to promote conscious awareness of unconscious, schematic functions. Such awareness may permit the individual to actively override some of the influence of the unconscious information processing.

Scripts or schema may be further organized into metascripts (Singer & Salovey, 1991) that reflect the individual's style of dealing with scripts. These superordinate schema may explain resilience because they influence how individuals reshape their schema when confronted with daily stress, unforeseen failure, or unexpected upheaval. Those with more complex schema may be more willing to replay negative information to explore a variety of alternative schema or scripts. More flexible schema involve the realization that current ideas that seem so clear and self-evident at the moment may change as a function of new experience. This process cannot be taught directly, but is a byproduct of experiencing changes in perception, understanding, and feeling in light of new information. Consequently, it is important to assess the style of synthesizing experience through the organization and reorganization of schema.

Analysis of responses elicited by projective techniques can be informed by models of schema development. Abelson (1976) assumes that schema begin as representations of single concrete examples and become more abstract. Rudimentary schema are compilations of single, concrete examples that are used to make snap judgments about seemingly similar instances. Stereotypic schema encompass the most representative features of events, persons, or groups. More abstract schema recognize inconsistencies between reality and the activated schema. The individual is aware of complexities and ambiguities rather than simply encoding schema-congruent information. Highly developed schema permit the individual to engage in complex information processing seemingly effortlessly and without awareness. The more abstract the schema, the more flexibly they can be applied because they include conditional and inferential concepts, abstract rules, and affective information. The progression from rudimentary to more complex schema is the product of experience and the degree of active, strategic effort brought to the structuring and organization of experience. Schema theory recognizes the role of affect, motivation, and biological

factors (genetics, temperament), as well as environmental influences (stress and supports) as they interact with cognitive processes. The shaping of schema, especially social scripts by culture, is also acknowledged and has been studied by anthropologists and linguists (Quinn & Holland, 1987).

Responses to projective methods may be evaluated according to aspects of schema described above. When telling stories to picture stimuli, some individuals superimpose associative elements without an organizational network; others impose stereotypes that may or may not precisely fit the stimulus configurations; still others draw creatively on various elements of their experiences to tell a cohesive story that captures the gist of the stimulus and complies with the instructions. If an individual is presented with a TAT picture that cannot be readily explained by a stereotypic script or an intact schema, then the respondent is called upon to actively construct a script to fit the situation. Similarly, when faced with inkblots that are rough approximations of real objects, the individual must remain flexible in applying schema to answer the question "What might this be." Those who rely on more rudimentary schema will make associations directly from discrete portions of the stimuli. However, those who apply more cohesive schema find a meaningful framework with which to integrate disparate or complex stimuli.

Two aspects of long-term memory structures (schema), procedural and declarative knowledge (Anderson, 1983; Kihlstrom, 1984), are relevant to the interpretation of projective responses. Procedural knowledge refers to unconscious processes or skills such as the structure of language, the organization of music, or other implicit rules that order information or perception. Declarative schematic structures entail the recall of factual information such as names, locations, and historical events. Responses to projective techniques reveal the implicit organization of knowledge as well as the content that enters awareness in response to the stimuli presented. Interpretive approaches can focus on the organization of schema by examining the structural aspects of responses and the sequence of ideas expressed, as well as by analyzing the content.

4.16.3.2 Unconscious Processing of Information

The idea that human behavior is based on a considerable storehouse of organized knowledge structures operating outside of conscious awareness is now widely accepted by cognitive

psychologists (Bargh, 1994). Since the 1970s, increasing evidence suggests that information not accessible to conscious awareness influences memory, perception, and thinking (Kihlstrom, 1990). Efforts to refine assumptions about unconscious processes continue across the various subfields of psychology and are not restricted to projective testing or even to clinical psychology.

Social cognitive psychology (social cognition) has become increasingly interested in unconscious thought processes of mental representations of situations, self, and others (schema) and in the interaction of cognition and affect (Wyer & Srull, 1994). Attitudes, expectations, or schema that are strong enough to be automatically activated have been described in terms of their "chronic accessibility" (Fazio, Sanbonmatsu, Powell, & Kardes, 1986). Chronically accessible schema may be cued by emotions arising in ambiguous situations, including reactions to projective stimuli. According to Westen (1993), "Research on chronic schema accessibility has confirmed the projective hypothesis that in ambiguous situations enduring interests, concerns, needs, and ways of experiencing reality are likely to be expressed" (p. 381). Bargh (1994) suggests that temporarily accessible schema due to a recent experience can mimic chronic accessibility and constitute a pitfall for studies of enduring cognitive structures. Indeed, this pitfall is common in the use of personality tests, be they questionnaire or projective devices. The distinction between state and trait variables acknowledges states as temporary conditions of mentality, mood, levels of arousal or drive, and these states may be related to temporarily evoked schema. One of the advantages of projective measures is the opportunity to examine both the structure and content of schema even if temporarily evoked. The content of the schema may be more susceptible to influence by such variations in experience than the structure. However, data are needed to support this contention.

A dramatic demonstration of the effect of a recent experience occurred one semester when the author's students administered assessment batteries to incarcerated youths (generally between the ages of 18 and 22) to determine educational strategies. A great number of TAT stories ended by the characters "talking about it and solving their problem." After some investigation, it was learned that a communication program had been recently instituted to teach the value of resolving conflict by talking things out. A close inspection of the stories themselves revealed that the schema about communication were not well developed but verbalized as a stock ending apart from other

components of the problem-solving process. For example, the problems were poorly understood, and characters' feelings, perspectives, and intentions were vaguely or inaccurately deciphered, if at all. The idea for talking about the problem became accessible to conscious awareness through training as declarative knowledge but without the understandings or procedural knowledge needed for its implementation. Conscious awareness of the benefits of communication was not combined with the implicit and largely unconscious understandings and affective reactions that promote such interactions. This example illustrates the importance of studying the inner logic and cohesiveness of schematic structures to discover how story content derived from recent experiences (including vicarious ones such as books or movies) is integrated with other aspects of the task demand such as accurately perceiving the stimuli and following instructions. Both declarative and procedural knowledge must be addressed to overcome the pitfall of attributing too much weight to content that is not meaningfully incorporated into guiding structures.

4.16.3.3 Inert Knowledge vs. Usable Knowledge

Inert knowledge has been described as information that the person knows, but does not apply unless explicitly cued or prompted to do so (Bransford, Franks, Vye, & Sherwood, 1989). The reason why a person fails to use relevant knowledge to solve particular problems is that this information does not spontaneously come to awareness in the context of that problem, or that the schema for applying the information are not sufficiently developed. The manner in which knowledge is encoded and organized in the first place determines its subsequent accessibility and applicability in other situations.

Effective learning in general requires active strategies to organize and recall information such as rehearsal or breaking down the task into manageable units. Torgesen (1977), who introduced the concept of the inactive learner, reported that learning-disabled students exhibit a learning style characterized by minimal planning and limited use of strategies. Individuals may process information from one domain more thoroughly than in others, or the inactive stance may be more pervasive. This inactive style is manifested by processing information exactly as it comes in (memorization) without organizing and restructuring it. In the absence of active, effortful, strategic processing of information, the knowledge may

be dormant in memory and not available for use unless externally prompted.

Different types of tasks are needed to differentiate knowledge that is used spontaneously to solve problems from knowledge that is available only when cued by the context. Everyday behavior requires spontaneous use of prior knowledge to make social judgments or other decisions; the possession of factual knowledge is far removed from problem solving in everyday encounters. The TAT stories of the incarcerated youths referred to earlier shows how knowledge may be stated without being embedded in a usable framework. Projective measures differ from typical cognitive tasks and self-report measures of personality in ways that have relevance for distinguishing between inert and useful knowledge. Rather than being asked to provide specific information about the self, the respondent is provided with a relatively unstructured task and asked to give an open-ended response that is evaluated by a professional. Likewise, solving verbally communicated problems such as those on intelligence tests does not resemble the conditions of everyday situations. Such problems tend to contain words that act as cues for accessing relevant knowledge, whereas everyday problem-solving requires the individual to size up relevant features of the situation to notice that there is a problem (Sternberg, 1985). Real problems are accompanied by emotions that shape the way the individual thinks about the problem. Individuals may verbalize several acceptable alternatives for handling a situation, if prompted. Nevertheless, when faced with real-life dilemmas, they may be unable to translate this knowledge into appropriate action. The projective tests reveal the spontaneous accessibility of the schema that guide the application of what comes to awareness.

4.16.3.4 Affective and Motivational Influences on Cognition

Most studies in this area focus on the way mood or emotion influences memory and judgment (Isen, 1984, 1987) and on unconscious processing of information (Uleman & Bargh, 1989). Epstein (1994) describes the impact of emotion on thinking and argues for the existence of two fundamentally different but interactive modes of information processing: a rational system and an experiential system driven by emotion. Thinking in the experiential system is shaped by emotion and tends to occur without deliberate effort or conscious awareness. The rational system is characterized by a more deliberate process of thinking and of

acquiring information such as through textbooks and direct teaching. Epstein explains that stories and anecdotes spice up an otherwise dry lecture because they increase emotional engagement and thereby appeal to the experiential system.

Emotions dramatically alter the thought process and influence behavior. Thinking is transformed by intense emotion in the direction of being categorical, personal, concrete, unreflective, and action oriented. The product of such thinking is likely to be considered as self-evident and not requiring proof (Epstein, 1994). Depressed affect is associated with preference for immediate gratification in place of delayed but more substantial rewards (Wertheim & Schwartz, 1983). Anxiety tends to bias the interpretation of ambiguous text (MacLeod & Cohen, 1993). Chronically experienced negative emotions appear to have a cumulative influence on the development of cognitive structures amenable to assessment with projective techniques. Children temperamentally prone to experience negative affect told stories to TAT stimuli that reflected less complex information processing and more reactive self-regulatory styles dominated by more immediate concerns than their less emotional peers (Bassan-Diamond, Teglassi, & Schmitt, 1995).

The following vignette shows how intricately problem solving is tied to emotions (*Readers Digest*, May, 1982, p. 79. Quoted from Jim Whitehead, quoted by Seymour Rosenberg in Spartanburg, SC *Herald*):

Two hikers were walking through the woods when they suddenly confronted a giant bear. Immediately, one of the men took off his boots, pulled out a pair of track shoes and began putting them on. "What are you doing?" cried his companion. "We can't outrun that bear, even with jogging shoes." "Who cares about the bear?" the first hiker replied. "All I have to worry about is outrunning you."

Now what if the two were husband and wife, parent and child, or very close friends? The emotional investment in the relationship would change the problem to how can "we" escape or even how can "you" reach safety. A parent may suggest that the child run for it. The automatic influence of affect in defining the range of acceptable solutions to problems is evident. Westen (1993) observes that cognitive psychologists do not focus sufficiently on the motivational and emotional factors of unconscious processing of information. Affect and motivation can enhance or disrupt the degree of organization and resourcefulness in the acquisition and subsequent application of knowledge structures.

4.16.3.5 Narrative Psychology and Thematic Apperception Techniques

The analysis of narrative has been a rich source of data in studying the schema or scripts that individuals bring to current understanding of experiences. The social-cognitive approach examines the memory system by analyzing the individual's style of processing and organizing information. Memories such as self-defining experiences (Moffit & Singer, 1994) or incidents of being rejected (Baumeister, Wotman, & Stillwell, 1993) are examined to extract abstract rules and cognitive principles. A study on script formulation (Demorest & Alexander, 1992) first asked participants to generate autobiographical memories. Scripts extracted from these memories were subsequently (one month later) compared with those derived from participants' invented stories in response to affective stimuli. The scripts drawn from fictional stories were similar to those derived from autobiographical memories. These findings suggest that scripted knowledge structures are superimposed on new affective stimuli.

Schank (1990) addressed the question of how stories exchanged during the course of social conversation are accessed in memory. To understand how a person is reminded of a story to tell at the right moment in a conversation requires breaking it down into themes, plans, goals, actions, and outcomes (Schank, 1990). These elements constitute a filing system for indexing the relevant components of experiences to be stored in memory and subsequently retrieved. A higher level organization of these concepts involves a prediction, a moral, or a lesson learned. When a lesson repeatedly occurs, it becomes a type of structure that exists apart from the specific stories from which the lesson arose.

For Schank, telling a good story at the right time is a hallmark of intelligence because it represents an understanding of what is being talked about now through its connection with the lessons that the listener has stored in memory. Recalling an appropriate story to tell and understanding how it relates to what is currently being spoken about depends on how the stories are catalogued in memory. It seems logical that individuals who engage in more active and reflective classification of experience for storage into memory are also more likely to be reminded of an appropriate story at the right moment in a conversation. This framework is directly applicable to the interpretation of stories told to TAT-like stimuli. Respondents are reminded of a story by the stimulus, but the particular story details are shaped by convictions or lessons learned from experience.

One way to validate constructs is through the congruence of models drawn from various lines of research. Arnold (1962) proposed that individuals express their basic attitudes or convictions in their TAT stories, and that these convictions constitute their motivational set. Like Schank, she focused on goals, plans, actions, and outcomes as the basic ingredients of the story import or lesson learned.

The rise of narrative psychology has the potential to enrich story-telling methods of personality assessment. Analysis of narratives such as those elicited by the TAT is considered useful for clarifying the degree of complexity and organization of information processing pertaining to the synthesis of experience (Carlson & Carlson, 1984). These cognitive and perceptual processes can also be assessed through the Rorschach technique. It has been suggested that the Rorschach should combine its emphasis on perception with a focus on cognitive representation (Blatt, 1992).

4.16.4 LEVELS OF PERSONALITY ASSESSED BY PROJECTIVE AND QUESTIONNAIRE METHODS

Projective methods measure different dimensions or levels of personality than self-reports. One of the clearest ways to understand the unique contribution of projective techniques is in the context of understanding these levels. McAdams (1995) proposed three conceptually distinct levels of understanding personality: level one—traits or stylistic, habitual tendencies; level two—developmental or motivational constructs such as goals, plans, and strivings; and level three—the life narrative or the evolving story that is internalized to provide broader meaning, purpose, and cohesiveness to specific experiences. McAdams does not describe relationships among the levels, though it seems logical to assume that they are not independent since they are different perspectives (or units) on the functional whole. For Klopfer (1981), a trait itself can be understood at three levels. First, as viewed by significant others or the public image (with or without awareness); second, as viewed by the individual or conscious self-concept; and third as manifested (with or without awareness) in behaviors such as responses to projective tests. This emphasis on the source of the data is consistent with a fundamental distinction between two levels of personality, based on the perspectives of the actor and observer (Hogan, 1987). Projective tests were better able to predict peer rankings on four personality traits (publicly observed image) than the individual's self-

report of those traits (McGreevy, 1962). Different techniques may be useful for elucidating different levels or dimensions of personality. However, a comprehensive understanding requires a framework for how the different levels relate to one another.

Measuring achievement motives through self-report of intentions, goals and reasons for actions (self-attributed) and inferences drawn from stories written to picture stimuli (implicit) are not equivalent (Koestner, Weinberger, & McClelland, 1991; McClelland, Koestner, & Weinberger, 1989). McClelland and colleagues propose that the general absence of significant correlations between measures of self-attributed motives derived through self-report, and implicit motives inferred from stories, cannot be ascribed to the worthlessness of projective measures nor to poorly designed questionnaires, but must be taken seriously as evidence that these are in fact different variables with the same name. Rather than seeking to remedy deficits in either type of measure, they suggest that it may be more fruitful to acknowledge them as reflecting two qualitatively different kinds of human motivation. They contend that implicit motives develop by intrinsic enjoyment generated when doing tasks or experiencing activities, and predict spontaneous goal-directed actions sustained over time even in the absence of specific social demands. However, self-attributed motives are built around explicit social incentives or demands of socializing others, and predict responses to situations structured to provide such incentives. The implicit achievement motivation does not give information about the area of life to which a person will direct efforts to succeed. Self-attributed motives, plans, and goals may express the person's conscious intentions but do not give information about the person's commitment and capacity to follow through. Under specific conditions, self-attributed motives act like implicit motives in that they energize, direct, and select behavior. However, the problem with predicting long-term behaviors from self-attributed motives is that the social incentives may not be salient enough to elicit the behavior. Implicit motives exert relatively greater influence because they drive activity that is inherently enjoyable even in the absence of specific social demands.

The acquisition of implicit and self-attributed motives such as proposed by McClelland and colleagues may relate, respectively, to the experiential and rational systems discussed earlier (Epstein, 1994). Learning in the experiential system takes place in an affective context where emotions and interests drive attention

and information processing. Additionally, the development of these two types of motives may involve different principles of learning (Meissner, 1974, 1981; Raynor & McFarlin, 1986; Sandler & Rosenblatt, 1962). One pertains to the acquisition of cognitive skills that are functional and enlarge the individual's repertoire of adaptive capacities through the teaching and reinforcement of these competencies. The other relates to the development of the inner world or the self-system that provides relative independence from external incentives. The advantage of developing this inner structure is to shift from dependence on external supports to increased reliance on self-regulatory mechanisms (Meissner, 1981). These regulatory functions increase the individual's capacity to master the environment, endure stress, and delay gratification of impulses.

The development of increasingly complex and differentiated self-regulatory structures or schema takes place at a level of organization that is different from that of acquiring behavioral patterns through reinforcement. Such development relies on the individual's integration of life experience to seek consistency within the self (Raynor & McFarlin, 1986). However, emotional, cognitive, and attentional limitations interfere with synthesis of life experiences and detract from the development of inner self-regulatory structures. For example, the domination of awareness by that which is immediately striking or concrete is not compatible with the development of long range interests, values, or aims. The primary affective impediment to the internalization of such self-regulatory values are the limitations in the capacity to sustain emotional involvement in interests that are relatively remote from personal needs (Shapiro, 1965).

Motivation can be understood in relation to two necessary ingredients: the goal or intention and the self-regulatory capacity to sustain goal directed action (Arnold, 1962; Kuhl & Beckmann, 1994). Thus, a stated goal or intention such as playing violin in a symphony orchestra may be difficult to implement and sustain, regardless of musical talent because of the rigorous independent practice required. Such sustained effort is facilitated by the enjoyment of and interest in the activity. Otherwise, careful programming of external incentives is needed.

Understanding the gaps between a person's stated goals or intentions and the individual's self-regulative resources is crucial to planning interventions. A 19-year-old college student was referred by her therapist for an evaluation because no meaningful discussions were taking place after six months of regular attendance of

sessions. Sandra was in therapy at the insistence of her family because "she had no direction in life." During the initial interview with Sandra, she described her erratic schedule of eating and sleeping, her general lethargy and boredom, as well as sporadic class attendance. She reported having similar problems in high school such as difficulty getting up, missing school once a week, not studying, and cheating to get by. She stated emphatically that she was not depressed.

When asked directly about drug use, Sandra reported significant involvement starting in the fifth grade. She did not discuss this issue with her previous therapist because "no one asked me." She described the therapist as not providing enough structure during the sessions so she just talked about meaningless topics. Sandra verbalized her desire to achieve and recognized that she was not accomplishing her goals. At the same time, she admitted that she found school work aversive.

As part of the evaluation, the TAT, the Wechsler Adult Intelligence Scale (WAIS-R), and Rorschach were administered among other measures. The first five TAT stories that she told are given below to illustrate the lack of connection between Sandra's goal setting and means toward their attainment. The emphasis is not on developing a comprehensive clinical picture but on presenting general conclusions about each story that are pertinent to Sandra's motivational schema:

Card 1. Don't like pictures. Not all of them . . . Umm . . . like story—like once upon a time. This is a boy . . . He's . . . OK, he . . . is it a violin? I don't know. Can I say two things? (Whatever you want). He's either frustrated because he can't play it or he's sad because he's not allowed to play it anymore. He wants to play it but for some reason he can't . . . (stares at picture) (TO) He becomes a great (did I say that was a violin?) violin player, and he's happy. (OK, I think you have the idea.)

Sandra is immediately put off by the task and vacillates in her interpretation of the picture as she does in her own intentions. The story content is appropriate to the stimulus, but Sandra cannot determine whether the boy is experiencing internal (frustration) or external (not allowed) barriers to achievement. In any case, the boy takes no heed of these problems and becomes a great violin player, seemingly without effort. This unrealistic connection among circumstances, intentions, actions, and outcomes suggests that Sandra's schema do not include a clear grasp of the obstacles to achievement nor of the steps needed to attain goals.

Card 2. God! These things are ... OK, there's ... (stares) ... This girl on her way to school, she's walking to school, and she sees this man working in the field, and that's his wife watching him. His wife's talking to him I guess, may be talking to him. He's working with his horse on the farm, and she's thinking that she doesn't want to look like that when she gets older (like the lady). She doesn't want to live on the farm. She feels sorry for that lady, for both, for the lady. So she goes to school to get an education so she's not like them.

The girl goes to school not because she really wants to learn but to avoid becoming what she fears. Sandra's schema about the importance of education does not include a positive affective quality nor the spontaneous accessibility of the effort involved. Sandra does not delve into the purposes and inner life of the characters but relies on the stimulus, on external appearances (doesn't want to look like the lady), and social convention (getting an education). Thus external structures and supports are important in guiding her reactions.

Card 3BM. Don't know what this is ... umm ... guess it's a girl. A boy or girl, either sleeping, crying, or ... umm ... (frustrated) don't even know what. (repeat directions) I guess she's crying because ... OK, this doesn't make sense though. (stares as if answer is in the card) I think they're car keys. No idea. This is dumb. She's crying because ... what is that thing is ... someone stole her car. She has no way home from this place. So then she calls the police and tells them that her car is missing. Eventually they find her car, and then she's happy. Can you tell me what that is? (It can look like different things to different people.)

Sandra initially vacillates before settling on a plot. She became frustrated, stated that the task was dumb, but responded to encouragement. This tendency to blame outside factors is consistent with her reliance on the environment to regulate her behavior. The outcome of getting the car back only deals with part of the problem. Having "no way home from this place" was not addressed. Sandra's incomplete processing of information (only dealing with part of the problem, leaving out details called for in the directions) and limited resources (no internal representation of possible support; indecision and disjointed approach to narrating the story) suggest why she cannot regulate her behavior.

Card 4. OK, ... this is a man and a woman who ... they either just had a fight, or he wants to go somewhere, and she doesn't want him to go. OK, she's trying to persuade him either not to go or to forgive her. And then, I guess eventually he does it, he forgives her or he doesn't—or he might go, or he'll stay. And they live happily ever after.

Again, Sandra cannot commit to one story line, and the characters' intentions or reasons for acting are not examined. As with the previous stories, she seems unable to develop her ideas clearly beyond the initial identification of the tension depicted in the card, suggesting dependence on external stimuli. Her vacillation about what's going on in the picture (noted in several picture cards) suggests that she feels uncertain about her judgment in situations that are somewhat ambiguous. As in previous stories, the ending is unrealistically positive as Sandra does not make appropriate connections among intentions, actions, and outcomes.

Card 5. OK, can you make up people who aren't there? This is a lady who is coming to tell her children it's time for their nap. So she sticks her head in and says it's dinner time, and all the kids come to dinner. Everybody eats dinner, and kids go back and play. In the library, den—I think it's supposed to be the library.

Sandra's query about introducing additional characters further points to her need for external guidance. The lady intends to tell the children that it's time to take their nap but ends up calling them to dinner. This twist in the story suggests that Sandra is not monitoring story details (also seen in other stories) and that she has similar difficulty keeping track of her own intentions.

Generally, Sandra is able to recognize the conflict or tension in the pictures but has difficulty sticking to one explanation. Sometimes, she does not monitor story details sufficiently, leaving gaps or contradictory details. She does not examine feelings, thoughts, or intentions and has difficulty connecting planful, realistic actions with outcomes. She cannot develop the themes that she introduces beyond a vague and Polyanna-like story. Sandra persists in staying in school because she believes that the way to ensure her future is to get an education but is not able to get invested in the process of learning. Sandra is floundering in the unstructured campus setting because she has not developed the inner structures (implicit motives) to regulate her attention and behavior.

When she received the results of the evaluation, Sandra easily identified with the example of someone who thinks an exercise program is needed and joins an aerobics class but does not attend for one reason or another (too tired, too cold). She agreed that even the thought of doing school work is aversive, and the key was in finding ways to side-step the aversive elements, possibly through social supports. She acknowledged that by joining a sorority, she would have peers to eat with on a regular basis, and the structure would help her keep a schedule and attend class. At first, she worked on seeking the external incentives that would motivate her behavior and through these initiatives began to feel she had some choices and control over her life. Subsequently, therapy sessions focused on ways to build these structures and prompts into her daily routines to gain increasing independence.

Sandra's schema as assessed with the TAT did not include connections among goals, plans, actions, and outcomes, and these resources were not available to guide her daily behavior. The therapeutic aim was to help Sandra deal with the factors that impeded the development of these connections. She could not have verbalized these impediments in an interview or self-report measure because she was not aware of them. Likewise, the more structured Wechsler scales did not reveal the sources of her difficulty. Therefore, sole reliance on such measures would have been insufficient.

4.16.5 SELF-REPORT VS. PROJECTIVE PERSONALITY TESTING

The foregoing discussion has already demonstrated that self-reports, even if candid and accurate, may predict behavior only when social incentives are operating or when the situation is structured to prompt the response. However, it is useful to contrast questionnaire and projective devices as direct and indirect methods of personality assessment (Vane & Guarnacia, 1989). Direct methods are comprised of observations of behavior, inventories, or interviews. The indirect method utilizes relatively unstructured stimuli such as the TAT, Rorschach, or drawings. In the latter, the information sought is not readily apparent and, therefore, difficult to manipulate. Vane and Guarnacia cite the following problems with the direct method: (i) limitations of self-knowledge preclude accurate self reports; (ii) test items are open to misinterpretation; (iii) real-life situations cannot be represented by pencil and paper items; (iv) desire of the respondent to manage impressions need to be overcome by

including scales to detect faking or response sets; (v) inventories yield information without clues for understanding the reason; and (vi) the relationship between traits on inventories and behaviors is complex.

Multiple choice or other highly constrained response formats do not give clues about the style of organizing the response. The respondent is asked to tell versus show. Thus, no evidence exists beyond self-report to judge response parameters despite availability of norms. Finally, responses to structured questions only tap conscious phenomena. Therefore, methods such as interviews or self-report inventories provide what the individual is capable of perceiving and is willing to share. However, responses to projective testing can be evaluated by analysis of various parameters.

Both direct and indirect sources of information are useful depending on the purpose. Self-reports are susceptible to conscious or unconscious distortion and reveal only one dimension of the personality. Nevertheless, it is important to obtain information that the respondent willingly discloses. Self-reports are easily scored and amenable to traditional methods for estimating reliability and validity. Projective methods require extensive training, are time consuming, and require more cumbersome methods to establish psychometric credibility. However, their purpose is not evident to the respondent, and they provide information that is different from other sources of data. The use of both direct and indirect methods is likely to yield the most accurate picture.

4.16.6 CONTRIBUTION OF PROJECTIVE TECHNIQUES TO DIAGNOSIS AND INTERVENTION

The process of responding to the projective task can be conceptualized as an attempt to impose meaning on the stimuli presented and to comply with the instructions given. Analysis of the response reveals the individual's schema or scripts, both in terms of the structural organization of experience and the content of awareness. Appreciation that the manner in which individuals acquire, organize, and represent knowledge into cognitive schema have a fundamental influence on the way they behave, guides the use of projective techniques. Rather than looking for one-to-one correspondence between test responses and inferences, such as searching for aggressive content in projective tests, it is more fruitful to explore processes that promote aggression in particular situations. The social-information-processing model of children's social behavior (Crick &

Dodge, 1994) details the social-information-processing steps that precede aggressive behavior in specific situations. According to this model, aggressive behavior can be understood in terms of how individuals encode social cues, how they interpret those cues, how they select and clarify goals, how they generate possible solutions to perceived problems or dilemmas, how they make decisions about the selection of responses, and how they execute and monitor the selected behavior. Information regarding each of these steps in social information processing can be gleaned from responses to projective techniques.

Thematic apperceptive methods can provide information about the attribution of intentions for others' reactions, the interpretation of social cues, the anticipated consequences of alternative actions, as well as about the cognitive flexibility and organization of the individual's meaning structures. Information about the execution of the behaviors is provided at two levels: how the characters follow through on intentions and resolve problems, and how the narrator develops and monitors the story. Any aggressive content that is expressed can then be better understood in light of the aforementioned processes.

Responses to the Rorschach also clarify social information processing that promotes aggression. For example, a preponderance of answers that simplifies the stimulus (e.g., high lambda) suggests that the individual may be involved in confrontational situations because of a tendency to make decisions without considering important cues (Exner, 1993). The connection between such responses to projective tasks and aggressive behavior is borne out by research suggesting that aggressive children selectively attend to and recall aggressive cues, partly because they respond prior to processing all of the available information (Dodge & Feldman, 1990).

A taxonomy of psychopathology comprised of a descriptive, atheoretical compilation of symptom categories (e.g., *Diagnostic and statistical manual of mental disorders*, 4th ed. [DSM-IV]) could be enhanced by a framework that relates these symptoms to basic structures of psychological organization. Within a diagnostic category such as autism, there exists a wide variability of impairment in functional capacities. Responses to projective techniques can range on a continuum from being flexibly adaptive to demonstrating various levels of impairment that can go beyond a specific diagnosis. The manner in which an individual uses the cues in the stimuli and responds in an organized way to the task demands is an index of the differentiation and organization of

previously developed schema. Furthermore, responses reveal whether the person must be prompted to respond or actively restructures knowledge for adaptive use in stressful or ambiguous situations.

Horowitz (1991) envisions a future *DSM* Axis II as incorporating three interconnected components: (i) recurrent maladaptive patterns of self-regard and interpersonal behavior; (ii) person schematic characteristics explaining that behavior—this would include the developmental level of self-organization; and (iii) style of self-regulation comprised of habitual control processes leading to coping and/or defense. An example would be how a person regulates emotionality by inhibiting or facilitating the activation of schema. Horowitz's conceptualization implicitly recognizes that any Axis I diagnosis, such as attention-deficit hyperactivity disorder relates to the development of inner schema. Difficulty regulating the attentional process may interfere with the active and strategic synthesis of knowledge gained through experience. Without organized schematic structures, the individual may not be able to govern behavior according to internalized rules and standards, a prominent characteristic of individuals with attentional deficits (Barkley, 1990). Responses to open-ended projective tasks represent the convergence of all of the individual's traits on the organization of experience, including compensatory strategies that could mitigate against risk factors.

A changing emphasis within psychoanalytic theory favors the view of psychopathology as representing an impairment in the formation of psychic structures that are analogous to the schema of social and cognitive psychology. This focus has spurred diagnostic efforts to evaluate the quality of these structures (self-system, object relations). Blatt (1991) assumed that the symptomatic manifestations of various forms of psychopathology are associated with different types of impairments of cognitive-affective structures in the representational world. For example, the growing consensus about the existence of two basic subtypes of depression (empty and guilty) is based on the phenomenology of experience associated with each. One type involves exaggerated preoccupation with issues of interpersonal relatedness, feelings of depletion, dependency, helplessness, or loss. The second type occurring later developmentally is associated with issues of self-definition, autonomy, guilt, and feelings of failure. The two types of depressed individuals change in various ways in the treatment process and are differentially responsive to different forms of therapy (Blatt et al., 1988). Wilson (1988) showed how each of these forms of

depression is demonstrated differently on the Rorschach and TAT protocols. For example, stories told to TAT pictures with single characters that describe the person as empty, worthless, or helpless without introducing other characters exemplify the first type of depression. Such stories suggest that the narrator is dependent on what is immediately evident in the surroundings without being able to use inner resources to represent the support needed. The vulnerability to depression of an individual lacking such resources may be moderated by a supportive environment.

The profession is in need of a diagnostic system that identifies variables useful for the treatment process (Leve, 1995). Such a system would emphasize characteristics that interact with methods of intervention and, therefore, would serve as guidelines for their selection. The variables that Leve tentatively proposed are grouped into three categories: cognitive, environmental, and social-emotional. The cognitive and social-emotional categories include the following characteristics, among others: adequacy of causal reasoning, social-emotional reality testing, moral reasoning, ability to form and maintain close relationships, ability to identify and express emotions, as well as internal or external locus of control. Leve (1995) reasoned that cause-effect thinking is essential to the success of psychodynamic therapies since developing insight is part of the process. Cognitive treatments, however, do not require as high a level of causal reasoning because the therapist is more actively training the client either by explaining explicitly the causal connections or bypassing explanations altogether through giving behavioral instructions. Likewise, different therapies require different degrees of social intimacy for success, although all therapies have an interpersonal component. The interface of psychological variables and therapy processes need to be refined through further research. It is evident, however, that projective techniques can provide important data about these types of variables.

A fundamental consequence of schematic processing is the perpetuation of the schema in the face of conflicting evidence. Distorted schema that bias attention and information processing to conform to the original conception are related to psychopathology (Horowitz, 1991). The assessment of schema can guide psychotherapy by identifying these unconscious schematic distortions (Blatt, 1991). Bringing to awareness different schema of self (e.g., actual self, ideal self, ought self, dreaded self) and other resulted in positive therapeutic changes in both psychodynamic and cognitive therapies (Singer & Salovey, 1991).

The clinician needs a flexible repertoire of assessment techniques including projective methods because the information that each one provides is not equivalent. No method supplants or takes priority over others but contributes its part to the functional whole. The contribution of projective techniques is best addressed by looking at their place among other measures. The content and structure of responses to projective techniques can be compared with information from self and other reports. The problem-solving strategies and organization of ideas as expressed in projective responses can be compared with how these processes are expressed in more structured tasks. Each source of data is conceptualized according to its contribution to the cohesive patterns that include different levels of personality and variability in functioning under different conditions. A synthesis of these multiple layers permits more precise insight about functioning in various contexts.

Projective techniques clarify the organization and structure of the inner world. Their use is warranted whenever inner life and complex self-regulatory functions are important considerations, particularly when respondents are lacking insight or would be motivated to distort self-report. The task itself involves the utilization of previously organized knowledge structures to provide a response that is not well rehearsed. The linkages among the various elements of the response promote an understanding of the individual's organizational framework. Projective techniques add a unique dimension to the assessment by revealing the respondent's strategies for accomplishing the task and, at the same time, showing the content and organization of ideas that occupy awareness.

4.16.7 PROJECTIVE TECHNIQUES AS PERFORMANCE MEASURES OF PERSONALITY

The projective hypothesis implies that every human action and reaction bears the characteristic features of individuality (Rappaport, Gill, & Schafer, 1968). Therefore, projection is basic to every response process rather than specific to a certain set of stimuli. Essentially, any test is a performance measure designed as an analog to a life task. Projective devices are simply stimuli with known qualities that are used to elicit samples of behaviors that correlate with other behaviors.

The question has been raised about what aspects of responses to the Rorschach stimuli reflect compliance to task demands and what elements entail projection. Exner (1989) suggested that most responses are simply best-fit

answers, and that projection on the Rorschach occurs only in responses that deviate from the norm or are elaborated beyond the stimulus field. This view assumes that projection on the Rorschach is an attribute of the response and not inherent in the task. Exner explains that projection is not encouraged by the instructions which simply call for answering "what might this be" nor by the blots which are not completely ambiguous. He points out that techniques such as the TAT force projection, by asking the respondent to develop a story that goes well beyond the stimulus provided. Yet, the story-telling task also makes problem-solving demands (Holt, 1961); the narratives are expected to be organized productions rather than fantasies or random associations. Just as identification of blot contours are expected to be guided by reality constraints, so do perceptions of emotions and relationships as well as the sequence of events in stories told to TAT stimuli.

Exner claims that the operations contributing to the formulation of responses to Rorschach stimuli such as scanning, encoding, classifying, refining, evaluating, discarding, and selecting are cognitive and not projective. However, in keeping with the projective hypothesis, individuals are expected to superimpose unique styles of organization on each of these cognitive operations. The distinction between problem-solving and projection is not an either/or, but a matter of relative emphasis. All projective tasks have problem-solving elements, and all responses are influenced by individual sets or internalized schema. Those who favor the projective viewpoint insist that such inner structures are fundamental to the response process. The conceptualization of the Rorschach as a perceptual-cognitive-behavioral task has been criticized on the grounds that this approach does not give sufficient weight to the role of stimuli from the inner world in the interpretation of stimuli from the external world (Willock, 1992).

As discussed earlier, projection is an ongoing process that comes to the fore if the situation is novel. In a routine, overlearned situation, individuals will respond with familiar, scripted knowledge. However, even a highly structured situation can be misinterpreted. When ideas on the more structured cognitive tasks are associative or disorganized, these cognitive difficulties will be magnified in responding to the less structured performance tests. One youngster who demonstrated serious problems in the application of rote knowledge in a relatively structured task also showed impaired thought process on the projective tests. To the Wechsler Intelligence Scale for Children (WISC III)

comprehension item, "what would you do if you saw thick smoke coming from your neighbor's home?" he responded: "Stop, drop, and roll." This response was an association cued by some of the words in the item without understanding the entire question. The fire was in the neighbor's home, yet the child produced a rote response as if he were personally experiencing the fire. He had rehearsed this scenario at school, but was unable to apply the concept even under relatively structured conditions.

The capacity to organize the inner world and deal effectively with distracting thoughts, emotions, or motives should translate into the ability to tolerate and to deal successfully with ambiguity, complexity, and apparent contradiction in a variety of situations. Therefore, the experience of ambiguity and complexity may stem not only from reality demands but also from the individual's emotions and motives (Blatt, Allison, & Feirstein, 1969). A framework is needed for conceptualizing how projective techniques such as the Rorschach and TAT shed light on cognitions in ways that differ from more structured tests. Differences and similarities in task demands of various projective techniques also need to be understood.

The response to the Rorschach task involves the matching of ambiguous stimuli with a memory trace (Exner, 1993). The goodness-of-fit or perceptual match between the blot contours and the object reported constitutes the basic element of reality testing. When telling a story about a picture portraying one or more people, what is demanded is not simply perceptual matching required by the Rorschach task (Beck, 1981) but experiential matching. The narrator searches for a relevant explanation for the picture, then marshalls possible details from memory to satisfy the instructions and meet criteria for an acceptable story. If the stimulus is highly stereotypic, then finding an exact story from memory would be sufficient to accomplish the task. However, if the narrator encounters a picture that cannot be explained by existing schema, then the individual actively constructs the response and fills in details from various schema. Schank (1990) suggests that what exists in memory is a database of partial stories or story elements rather than whole ones. When telling a story to a pictured scene, the narrator draws on this database to construct a set of events and inner processes (characters' thoughts and feelings) that captures the gist of the stimulus.

All tasks are on a continuum of how much organization is required in making meaning of the stimulus and developing the response. The more open-ended the task, the greater the demand to construct the response actively

rather than to respond in rote fashion to the stimuli. Therefore, projective techniques are particularly useful when adjustment to routine circumstances is good, but the individual is experiencing problems in less structured situations. Task by task analysis of the assessment battery permits the identification of the specific processes or competencies required for each task and the life situations to which these competencies can generalize.

A wide range of human ability lies outside the domain of standard cognitive tests. It is these competencies in the broad sense that are within the purview of projective testing. The primary advantage in regarding projective methods as performance measures of personality is that various tasks and measures can be differentiated in their demands. Generalization from one test performance to another and to real life performance then can be based on similarity of the demands. If projective techniques are treated as performance measures of personality, then it is possible to delineate task requirements, norms, and expectations about the product as seen with the Rorschach. In addition, it is possible to delineate linkages of that product with the acquisition of prior knowledge structures and with the application of these resources to similar problem-solving situations.

The term "*performance test of personality*" is proposed because it captures two key dimensions of projective techniques: (i) the problem-solving aspect of meeting specified performance expectations such as form quality on the Rorschach; and (ii) the organization of inner resources or schema brought to bear on the production. This dual approach to understanding projective techniques is shown in Table 1.

The distinction between self-report and projective measures has been described earlier. Self-reports ask the respondent to tell about the self, whereas projective measures require the performance of a task from which psychological processes are inferred. If we want to know how a person solves certain types of math problems, rather than ask the individual to report on how well he or she can solve linear algebraic equations, we would present some sample problems. Similarly, if we want to assess personality, we could obtain information through self-report by asking questions such as how flexible are you? How do you interpret cues in an unstructured situation? In contrast, we could request that the individual perform a task that demonstrates how various psychological processes are applied to meet the problem-solving demand.

Tests such as the Wechsler Scales are generally seen as measures of cognition, whereas the TAT and Rorschach techniques are tasks

that are thought to reveal personality. This distinction is not entirely a function of the task but includes how the performance is evaluated. Human figure drawings, for example, have been used to estimate both cognitive level and personality functioning. The essential characteristics of tasks that measure personality, in contrast to cognition, are described below.

4.16.7.1 Total Personality

As do life situations, tasks vary in terms of the amount of structure, cues, or prompts provided for responding. As the expression goes, "It's not what you know but when you know it that matters." To be useful, knowledge structures must be spontaneously available when circumstances warrant. In the psychological test battery, personality performance tasks provide minimal guidance to meet the problem-solving demands and, thereby, permit the assessment of spontaneous strategies.

Generally, cognitive measures such as tasks on Wechsler scales do not reproduce the conditions encountered in everyday situations. Sternberg (1985) distinguished between analytic intelligence that is assessed by more structured tasks and practical intelligence. Analytic problems provide all necessary information and have a single correct solution separate from the emotional and social context. Real life problems are less clearly defined and typically call for information seeking. Note the following question on the WAIS-R: "In a movie theater, you are the first person to notice smoke and fire. What should you do?" The respondent is cued that the expectation is to do something. If this were a real situation, the person would not know that he or she was the first to see the smoke and fire. The person may not take responsibility with others around and, therefore, delay action. The individual may consider specifics of the context such as proximity to the fire, its size, or the availability of a fire extinguisher. However, the individual's emotional reaction may disrupt thinking. The actual response would involve the total personality, not merely the intellectual component.

The limitations on generalizing from responses to structured tasks to real life conditions are evident when we consider the attributes that these measures do not assess. These include being aware that a problem exists or that a task needs to be done; setting priorities and planning toward their implementation; pacing and self-monitoring; seeking or utilizing feedback; sustaining long term interest in independent activities; taking necessary risks; organizational skills; and recognizing and

Table 1 Task analysis of performance measures of personality.

<i>Technique</i>	<i>Ideal performance</i>	<i>Competencies needed</i>	<i>Inner life (how experience is organized and internally represented)</i>
Drawing	Symmetry, coherence, realism, match with instructions	Plan execution of drawings within allotted space and limits on drawing ability; organize details with context, handle frustration; investment in the task	Reality testing and quality of thought process; use of schema and organizational strategies; specific preoccupations or concerns
Narrative	Story matches picture and incorporates instructions, appropriate transitions and cause-effect connections; synthesis of various dimensions of experience within individual characters such as integration of inner life with external circumstances; balance among views and needs of all characters depicted; balance between excessive detail and vagueness; cohesiveness among intentions, actions, and outcomes; appropriate time perspectives	Draw on personal experience to construct a story that adequately explains the picture; modulate affects and recognize tensions depicted in the stimulus; organize, plan, and monitor details of story for cohesiveness and inner logic and for compliance with instructions; direct attention from one aspect of experience such as feelings to actions and outcomes; initiative to transcend the pictured cues to describe intentions and realistic, purposeful actions to resolve tensions; bring general abstract principles to bear on the task for optimal integration of multiple dimensions of the stimulus and coordination of inner and outer aspects of experience (thought, feeling, action, and outcome); investment in the task	Reality testing and quality of thought process; understanding social cues; strategies to organize experiences; specific preoccupations and concerns; nature of inter- and intrapersonal schema or scripts brought to interpretation of experiences and picture stimuli (experiential matching)
Inkblot	Accuracy and specificity in matching percept to form; organizing various components of the blots; balance among form and other determinants as well as between precision and vagueness in form definitions; logical and responsive communication during inquiry	Adequate investment in responding; realistic perception and organization of stimulus components; understanding of hierarchical relationships implicit in balancing form with other determinants; comfort with matching precision of percepts to relative imprecision of blots; confidence with ambiguous stimuli; systematic approach to the task	Reality testing; strategic processing of information and connecting new input with previously acquired sets or schema (perceptual matching)

Note: The task is somewhat different for each specific set of instructions and stimuli (e.g., draw-a-person in the rain vs. draw-a-person). Similarly, each Rorschach card and or TAT picture presents a unique stimulus configuration.

responding appropriately to subtle interpersonal cues. These attributes are traditionally viewed as pertaining to the personality domain and reflect the interplay of cognitive and affective processes. A response to the TAT involves the interpretation of the emotions and tensions depicted to recognize a problem, whereas a verbally communicated question or problem often contains cues to facilitate the response, and such cues may not exist in a real-life version of the scenario. Likewise, the Rorschach presents the respondent with options for interpreting the blot contours, organizing the percepts, and communicating them to the examiner.

4.16.7.2 Many Correct Solutions

Just as real-life dilemmas are amenable to diverse resolutions, personality performance measures can be approached in several ways. Therefore, many correct solutions are possible. Responses can differ according to the individual's interpretation of the stimuli and organization of the response. Variability, rather than uniformity, is the expectation with personality performance tasks. Rather than imposing specific criteria for correct responses, such tasks set a general expectation to deal with the stimulus realistically, follow instructions, and logically organize the response. In contrast, most cognitive tasks have one correct solution regardless of whether they call for rote knowledge or a more flexible application of prior knowledge. These distinctions between cognitive and personality performance tests are not absolute; various cognitive tasks share some common processes with personality performance measures. Reading comprehension, for instance, bears a resemblance to personality performance measures when previous sets influence the understanding of the text. There can be little argument about the facts presented in a passage, but individuals can justifiably differ to some extent on their inferences based on prior learning.

4.16.7.3 Generalize to Different Criteria

The distinction between typical vs. maximal performance (Cronbach, 1970) has been used to differentiate between personality and ability measures. This dichotomy contrasts responses to immediate cues and external motivating structures from performance that is regulated and maintained by the individual's long-term investment and initiative. A structured situation such as an achievement or cognitive test would generally tap maximum performance, but such performance generalizes only to similarly struc-

tured situations. Because most real life tasks provide less structure, measures of typical performance are also needed in a comprehensive assessment. Correlations are low between measures tapping maximal and typical performance (Sackett, Zedeck, & Fogli, 1988). Therefore, each provides unique information.

The following conditions were suggested by Sackett and colleagues as generally yielding estimates of maximum performance: (i) there is a heightened level of effort and attention because the task is seen as important; (ii) expectations and performance standards are clear; and (iii) the observation takes place over a relatively short time where the individual can exhibit an uncharacteristic spurt of effort that could not be sustained over the long haul. In contrast, the characteristics of measures of typical performance are as follows: (i) individuals are unaware that they are being observed or evaluated, so they are not trying deliberately to perform to the best of their ability; (ii) performance is monitored over a long period of time; (iii) the performance tasks require skills that have to be learned through continuous past efforts—if the task is highly complex, the individual has to bring a great deal of past learning (typical performance) to the current effort; and (iv) performance guidelines are not clear and, therefore, individuals impose their characteristic way of organizing and dealing with the situation.

Performance measures of personality meet several criteria for assessing typical performance. Prior knowledge (schema) is superimposed on current task demands; individuals are unaware of what aspects of their performance are evaluated; and in the absence of structure, they are required to organize and plan their response according to their typical mode of functioning. Understanding different requirements of different tasks helps explain variations in the assessment of similar constructs with different methods. Johnston and Holzman (1979) designated indices of thought disorder in the WAIS and the Rorschach which they tested with a sample of schizophrenics, their parents, and controls. The IQ score and the Thought Disorder Index (TDI) derived from the WAIS were negatively correlated such that the higher the IQ score, the lower the TDI. In contrast, IQ scores and the TDI derived from the Rorschach were uncorrelated. The authors conclude that the tasks make different demands. The WAIS calls for "habitual reactions and the social frame of reference is clear" (p. 61). The social expectations on the Rorschach are less obvious. Therefore, the task sets different requirements and has different implications for the assessment of impaired thought process.

In their sample of schizophrenics, those who were more intelligent were able to limit expression of disordered thinking on the WAIS more easily than on the Rorschach, in part, because WAIS items can be answered with overlearned responses. The feeling of ambiguity versus security determines efficiency of functioning on problem-solving tasks and in real life. Individuals who exhibit thought problems only on unstructured tasks may be more capable of functioning with the supports of structures and clear social cues.

A criticism of laboratory research has been that results do not generalize to nonlaboratory situations (Fromkin & Streufert, 1976; Snow, 1974). This is the case because precise linkages between processes engaged by laboratory activities and other life domains are not specified. Likewise, predictions from test performance to real life adjustment can be accurate only if they are functionally similar. By understanding patterns of strength and weakness across various tasks, the professional can point to areas where the individual can and cannot function adaptively.

4.16.7.4 Differences in Conditions of Learning and Performance

Responses to performance measures of personality, as most tasks, require previously organized knowledge and strategies. However, we distinguish between learning that is promoted by direct teaching (e.g., lecture, textbook) from learning mediated by the individual's synthesis of experience (Epstein, 1994). Personality performance tasks are guided more by self-regulated learning than by formal education. Again, this distinction in the conditions of learning does not apply in a dichotomous fashion to personality versus cognitive measures. For example, an individual's general fund of information is a joint function of direct teaching and the individual's interest and active, effortful processing of information. Depending on the individual, knowledge is acquired through some combination of direct teaching and self-regulated synthesis of experience.

Task requirements also differ on the basis of the conditions of performance. These conditions pertain to the spontaneous versus cued accessibility of prior knowledge and degree of organization required to produce the response. Cognitive measures that assess general fund of information often elicit previously acquired knowledge in piecemeal fashion (highly structured). Personality performance measures such as the TAT and Rorschach set conditions of performance that demand spontaneous accessibility to prior knowledge to interpret the

stimulus and formulate the response according to the directions. The more open-ended the response, the greater the need for self-regulated strategies to organize the product.

Performance measures of personality maximize the imprint of organization so that the principles by which experiences are structured and the inner organization of the personality are revealed. A comprehensive assessment battery provides tasks that vary on the continuum of structure provided. Given such variation, it is possible to relate competencies required in each task to performance in life situations requiring similar competencies. This linkage is accomplished by understanding how processes exhibited during test performance are carried over to everyday functioning in various situations and incentive conditions.

4.16.8 SPECIFIC PROJECTIVE TECHNIQUES

Three basic aspects of projective techniques have been recognized (Rabin, 1981). First, the task entails the presentation of an ambiguous set of stimuli and a request to give an open-ended response. The Rorschach and TAT techniques include both of these stimulus and response attributes. Therefore, the focus of interpretation is on the perception of the stimuli presented and the organization of the response. Drawing techniques also demand open-ended responses but usually do not provide a stimulus. Second, the response is shaped by processes that are outside of conscious awareness. An important factor making the response less amenable to conscious manipulation (faking) is that the respondent does not comprehend the meaning of the answers given. Third is the complexity of the interpretive process. Each of these components is briefly addressed below.

4.16.8.1 Stimulus

In general, projective techniques present stimuli that are amenable to various interpretations, and instructions that can be addressed in a variety of ways. The degree of ambiguity has been a prime consideration, although other features of stimuli are also important determinants of the response. A systematic accounting of how the respondent uses stimuli is the essence of the Rorschach technique and must be given greater weight in thematic apperceptive methods (Henry, 1956; Teglassi, 1993).

4.16.8.2 Response

All projective tasks require the individual to draw on internal images, ideas, and relationships to create a response. The respondent must

dredge forth past experiences, direct or vicarious, and organize them to meet the task demands. The greater the stimulus ambiguity and the more open-ended the response, the greater the reliance on the organizational structures of the personality rather than on rote knowledge. Yet, the stimulus must have sufficient structure to permit evaluation of the plausibility of the respondent's interpretation. The projective task demand is an analogue of other unstructured tasks and situations where available cues are subject to interpretation. The manner in which the individual interprets the stimuli and organizes the response shows how they will respond under similar conditions (Bellak, 1975, 1993).

Responses to projective tasks involve complex, interrelated processes that have conscious and unconscious components. These include the interplay of cognition-emotion-action tendencies that coordinate perceptions of the outward world with experience of the inner. Projective techniques may reveal aspects of emotion, motivation, and cognition that a person may not wish to expose. These unconscious aspects of responses may relate to issues of invasion of privacy. Faking, malingering, or defensiveness are problematic for any form of assessment but are assumed to be less so with projective techniques. However, the issue of fakability of projective tests has been inadequately investigated (Rogers, 1997).

4.16.8.3 Task Demand

The stimuli presented together with the instructions set the task demands (Teglasi, 1993). Projective techniques impose task demands that cannot be met with a simple response such as a request for specific information. They require respondents to apply what they know to produce a story or a drawing or to identify an object that may fit the ambiguous contours of an inkblot. Various projective tasks have features in common yet differ in important ways. Projective methods have been understood as problem-solving tasks with designated performance expectations as well as measures of personality with emphasis on individual variation.

4.16.8.4 Interpretation

The interpretive task is complex even when the scoring categories and interpretive guidelines are straightforward as in the Comprehensive System for the Rorschach. Despite the relatively clear guidelines and availability of norms (not to mention computerized reports), the Rorschach requires the examiner's trained inference along with the more objective coding.

Interpretation is in keeping with the clinician's theoretical framework and understanding of the task demand. Therefore, the interpretation of responses to a projective test cannot be more satisfactory than the adequacy of the theory informing the interpretation of the evidence and the examiner's skill in evaluating that evidence.

The complexity of the interpreter's job is maintained because conclusions rest on the understanding of meaningful patterns rather than isolated response elements. Personality performance tests yield products where the whole is more than the sum of the parts. The evaluation of the response must account for the organization and cohesiveness of the different components. Therefore, various units abstracted from the whole cannot be treated in a piecemeal fashion. The more open-ended the response, the more amenable to separate analysis of structure, form, and style. Although these aspects of the response can be conceptually separated, their interpretive value lies in their relationship to the task demand and to each other. Content, for example, cannot be properly understood apart from the manner in which it is organized. Finally, the examiner must be aware of connections between empirical findings and theory to avoid speculation. Yet, empirical support of theory is not always conceptualized in ways that are useful in making decisions about one individual.

The sign approach attempts to provide empirical evidence for interpretations through the identification of features that occur most frequently in specified clinical populations. However, using a list of signs in an atheoretical cook book fashion is inadequate because a particular sign derives its meaning from the context of other responses. For example, in a TAT protocol, a stereotyped approach or meticulous listing of details in story telling may be viewed as resistance or as representing the respondent's best efforts. Likewise, concern with minutiae that most respondents disregard may be viewed as an index of hypervigilance or of concrete functioning. The appropriate interpretation depends on the pattern of responses within the story telling task and across tests in the battery.

Inferences are drawn from the clients' behavior during the evaluation as well as from the product. Any changes in demeanor or emotional reactions in response to the various tasks are noted as are spontaneous comments, time elapsed, expressions of uncertainty about performance, or attempts to seek structure. Aspects of the response process are also considered. These refer to the manner of working, compliance with instructions, sequence of ideas (e.g., planful, trial and error,

organized, haphazard), as well as to dysfluencies, pauses, or hesitations. The product itself is amenable to analysis in relation to the process of the response, content, and structure. The formal or structural aspects of the product, of course, reflect the response process.

Next is a brief review of three major projective techniques in terms of the following elements: (i) task demands that include the stimuli and instructions; (ii) response process; and (iii) general interpretive approaches and issues.

4.16.9 THEMATIC APPERCEPTIVE TECHNIQUES

Apperception tests generally use pictures and standard instructions to elicit stories. The TAT (Morgan & Murray, 1935) is the most popular of these. Although introduced as a method to assess a particular theory of personality (Murray, 1938), this technique did not remain wedded to the theory of personality from which it sprang. The thematic approach described in the standard manual (Murray, 1943) was amenable to use with a wide range of scoring methods and has been adapted for clinical and research purposes. Spin-offs of thematic apperceptive approaches have used different picture stimuli and diverse scoring approaches. A variety of nonclinical coding systems for content analysis of the TAT appears in an edited volume by Smith (1992a). More clinically oriented systems also abound (e.g., Bellak, 1975; Cramer, 1996; Henry, 1956; Karon, 1981; Rappaport, Gill, & Schafer, 1968; Teglasi, 1993; Tomkins, 1947; Westen, 1991; Wyatt, 1947). Typically, clinicians have used the TAT by applying broad units of inference in an idiographic, qualitative manner. In contrast, researchers have scored TAT responses for various personality characteristics using specific and narrow scoring criteria. As of this date, there is no widely agreed upon scoring system for the clinical use of thematic methods. However, it is generally acknowledged that classification into diagnostic categories is not a chief purpose of the TAT.

4.16.9.1 Response

When given TAT-like pictures, the scene sets the topic. The story telling directions call for a description of what is happening in the picture, what happened before, what people are thinking, how they're feeling, and how everything turns out at the end. These instructions require the narrator to attribute thoughts, feelings, and motives to the characters depicted and embed

these inner aspects of experience into an appropriate time frame and sequence of events that coordinate inner life with appropriate actions and outcomes.

4.16.9.2 Stimuli

Picture stimuli play a major role in determining the story content (Kenny, 1964; Murstein, 1965), and reviews of frequency of themes specific to TAT stimuli are available (Bellak, 1975; Henry, 1956; Holt, 1978; Murstein, 1968; Stein, 1955). Ambiguity was a primary consideration in designing the TAT stimuli (Murray, 1938). However, the degree of ambiguity varies within the set of TAT cards. The manner of defining ambiguity has ranged from judges' estimates of the number of interpretations that can apply to each card (Kenny & Bijou, 1953) to actual degree of variability in responses (Campus, 1976). Many of the TAT pictures clearly show who the characters are, what they are doing, and the emotions they are experiencing (Murstein, 1965). Yet, they permit a great deal of variation in the style of expressing similar themes. While stimulus ambiguity is essential to the projective hypothesis, other issues pertinent to stimuli such as similarity of main character(s) to the narrator, emotional tone, complexity, and latent meaning have been researched (see review by Teglasi, 1993).

Broad conclusions about the advantages of various degrees of stimulus ambiguity cannot be drawn without specifying the nature of the scoring system, population, and purpose. Low ambiguity has been favored in the assessment of specific motives (Singer, 1981). However, cards with high ambiguity have been preferred to measure the relative strength of two motives (Atkinson, 1992). When studying hostility, it was found that aggressive intent was most effectively measured by a picture with low relevance for hostility, whereas guilt over hostility was best measured with a picture of high relevance for hostility (Saltz & Epstein, 1963). The authors concluded that pictures with low relevance for unacceptable behavior such as hostility measure drive toward its expression, and those with high relevance measure inhibition or guilt about its expression.

Three degrees of ambiguity within the TAT set were estimated by a group of judges—high, medium, and low (Kenny & Bijou, 1953). The richness of personality content obtained varied as a function of card ambiguity. Cards in the medium ambiguous set yielded stories with the most personality information. The authors discussed two dimensions of ambiguity in card stimuli: (i) the number of cue constellations

available to guide the response; and (ii) the definitiveness of the cues available. Sets of stimuli with graded levels of ambiguity permit the evaluation of responses to situations along a continuum of structure.

Studies of stimulus ambiguity are inconclusive because investigators compare various degrees of ambiguity that are not on the same points on the ambiguity dimension. Furthermore, conclusions about stimulus variables have generally been drawn regarding their adequacy to elicit content germane to the assessment of specific motives or relatively narrow psychological processes rather than their effectiveness to assess formal qualities of organization that reflect the respondent's capacity to deal with the task demands.

For clinical purposes, stimuli must permit the simultaneous evaluation of multiple psychological processes and analysis of both structural and content properties. Some clinicians argue that stimuli are relatively unimportant as long as they are not highly structured (Arnold, 1962; Karon, 1981). Criteria for picture selection within the TAT set are available (Bellak, 1975; Birney, 1958; Henry, 1956; Teglassi, 1993). Haynes and Peltier's (1985) survey of clinicians in juvenile forensic settings indicated that the mean number of cards used was 10.25. The most frequently administered TAT cards were similar to those deemed as most productive in previous studies (Cooper, 1981; Hartman, 1970). The TAT pictures have been criticized for being out of date in relation to clothes and hairstyle (Henry, 1956; Murstein, 1968) and for their predominantly negative tone (Ritzler, Sharkey, & Chudy, 1980). These concerns are mitigated by findings that the stimuli effectively permit expression of convictions and conceptualizations of affect. The negative tone presents unfinished business or a dilemma to be resolved and provides the opportunity to observe how the respondent interprets the tensions depicted in the picture. Furthermore, the negative scenes make it possible to observe how the narrator moves beyond the sadness or conflict to a more positive resolution by noting appropriateness of transitions between the negative state of affairs and positive outcomes. It has been suggested that pictures representing relatively universal social situations that most people encounter in their lives are suitable across various age and subcultural groups (Veroff, 1992).

4.16.9.3 Interpretation

Karon (1981) suggests that only a complex scoring system can be clinically useful. At the same time, "an adequate scoring system is so

complex that no one has time to score the protocol" (pp. 94–95). Simple scoring systems are inadequate because they "throw away most of the information in the process of scoring, and hence turn out not to be clinically useful" (p. 95). This point is well taken. A multiplicity of thematic methods are available that provide carefully described scoring systems to assess relatively narrow aspects of personality. Although these systems have demonstrated a high degree of accuracy and reliability, their narrowness limits their clinical utility. Karon's remedy is to rely on clinical judgment in a sentence by sentence interpretation of the protocol, much like a clinical interpretation of an open-ended interview. This approach is compatible with the view that projective techniques are not tests, but are clinical tools that rely on the skill of the practitioner (Anastasi, 1976). An alternative is to agree on key psychological processes and guidelines for their measurement that may be time consuming to learn but less cumbersome to use once mastered.

Historically, there have been two ways to designate units of analysis for TAT stories: product centered versus narrator (person) centered. The former method focuses on qualities of stories that differentiate between groups; the latter approach emphasizes the psychological constructs to which attributes of stories relate. The narrator-centered approach permits the hierarchical organization of elements identified in stories in reference to broader features of the personality. For example, Rappaport, Gill, and Schafer (1968) organized story qualities according to affective lability and looked for content shaped primarily by affective responses to the picture. Holt (1958) looked at clarity of thought (vagueness, overgeneralization, disjointedness of organization) and emotional inappropriateness (arbitrary turn of events, forced endings). With such a focus on psychological processes, patterns of story elements can be organized in relation to relevant constructs.

Holt and Luborsky also emphasized adequacy of hero. This variable, however, is not a psychological process but a quality of the story considered important because it correlated with supervisor's ratings of overall competence during psychiatric residency training. Adequacy of the hero was subsequently viewed by Bellak (1975) as an index of the narrator's competency. However, this interpretation may not hold if other characters depicted in the stimulus or introduced in the story are helpless. This conclusion is also problematic if the narrator has not accurately incorporated the cues provided in the stimulus. A narrator-centered

approach to this variable would focus on various psychological processes indicative of the story teller's competency. Three points are relevant here. First, no part of the story can be interpreted separately from the others. Second, the emphasis must always be on the narrator rather than the narrative. Third, the demands set by stimulus properties must be considered.

Emphasis on qualities of the narrator rather than of the product permits consideration that these qualities can be expressed in different ways. Therefore, the interpreter does not seek one-to-one correspondences but focuses on the fit between the pattern of responses and the psychological processes to which they pertain. For example, the assessment of a construct such as cognitive integration with thematic apperceptive methods may be accomplished by exploring the manner of interpreting stimuli or the manner of coordinating different dimensions of experience such as thoughts, feelings, and actions. More complex units, such as self-regulation, subsume cognitive integration as well as affective-motivational processes (Teglassi, 1993). Such a focus on the psychological processes of the narrator avoids the sign approach and promotes an understanding of the constructs being measured. This occurs by organizing story qualities in terms of a clear conceptualization of their relationship to relevant psychological variables. Conceptualizations of the psychological variables can undergo continuous refinement in keeping with research across various subdisciplines.

Form and content in the TAT have been distinguished (Henry, 1956; Holt, 1958; Teglassi, 1993). Formal features focus on generalized properties of content such as the organization and coherence of the response. These formal units of analysis refer to how the details of the content relate to each other, to the evoking stimulus, and to the directions given. For example, inferences can be made from TAT stories about striving for long range goals (independently of the specific goals or concerns) based on connections among characters' purposes, actions, and anticipated outcomes. Internal attribution of feelings is a formal quality because it need not refer to any specific feeling. These formal qualities of the story provide information about the cohesiveness and reality base of the narrator's schema rather than about specific concerns. They are akin to procedural knowledge within schema theory and should be evident across stimulus cards within a task and even across various projective tasks. Formal features of stories lead to inferences about the structure of personality such as cognitive integration, affect maturity, or self-regulation. Given that picture stimuli evoke

specific content, it is these generalized or formal properties of content that can reliably reveal clinically important psychological processes.

A focus on the formal elements involves the following general principles of TAT interpretation.

(i) An organized narrative, not a patchwork of associations, is expected. The story is, therefore, evaluated in terms of how the narrator meets problem-solving task demands. The individual's problem-solving approach can be examined from two vantage points. The first focuses on how the story is told in relation to the stimulus and instructions. Accordingly, the narrative is evaluated in terms of accuracy in capturing the tensions depicted in the stimuli, compliance with instructions, and the logical and realistic unfolding of events. The second emphasizes how the characters are described in reference to the stimulus cues, the cohesiveness of inner states, actions and outcomes, as well as the manner in which characters define and resolve the dilemmas set before them in the stimulus.

(ii) Units of inference based on interconnections of elements such as the links between causes and effects or actions and outcomes pertain to formal characteristics of content. The content is not taken at face value. Instead, units of inference are based on understanding of psychological processes and task demands. The instructions call for inclusion of various levels or dimensions of experience such as thoughts, feelings, and actions. These human tendencies to think, feel, perceive, and act are interrelated, and interpretive meaning is derived from an understanding of their patterns. However, the narrator's understanding and the clinician's framework for conceptualizing such relationships must be clearly distinct (Cramer, 1996). The professional determines the implications of the narrator's understanding of the linkages among sequences of events, thoughts, feelings, behaviors, and outcomes.

(iii) The interpretation seeks patterns that elucidate the schematic structure connecting the various story elements (e.g., the manner in which self and other are differentiated and expectations about causal sequences). These schema are products of the synthesis of past experience and provide the templates for organizing current experiences.

4.16.10 RORSCHACH TECHNIQUE

The Rorschach has achieved wide acceptance due in large part to the development of the comprehensive system (Exner, 1993), an integration of previously established methods. The

availability of normative data for discrete coding variables establishes psychometric credibility. In addition, the comprehensive system provides interpretive strategies for synthesizing the complex data and encourages the interpreter to move back and forth between the formal data of the structural summary and the content and language of the response.

The Rorschach administration is conducted in two phases. First is the association phase during which the respondent is shown each of 10 inkblots and asked "What might this be?" The inquiry phase begins after the respondent reports what is seen on each of the cards. The examiner guides the respondent to clarify the determinants of each response to permit accurate coding. Although the Rorschach is described as a cognitive-perceptual task (Beck, 1981), what is interpreted is what the respondent verbalized. Hence, the examiner must be well trained in coding and administrative procedures to avoid the many potential pitfalls of conducting an inquiry and to maintain the delicate balance between seeking sufficient clarification and promoting response sets by pressing too far.

4.16.10.1 Stimulus Features and Response Parameters

The structural approach to interpreting responses to Rorschach cards creates a close link between the stimulus qualities and parameters of the responses that are coded. Stimuli vary in the degree to which they are solid or broken, colorful or achromatic. They also display variation in shading, empty (white) spaces, and contours that are sufficiently familiar to evoke popular responses. Reported perceptions are coded according to patterns of attending to stimulus qualities including choices of blot areas and use of various features of the stimuli such as form, color, shading, hue, white space, card symmetry, or any combination. Also important are the accuracy of the match between the reported percept and selected blot contours and the relationships among the percepts (organization). Each code assigned to a response represents a particular psychological process. However, specific response elements such as choice of blot areas are interpreted in relation to other variables such as degree of organization in relating various blot areas. A response involving the synthesis of parts into larger units reflects more complex thinking than a vague, undifferentiated holistic perception of the card.

Perception of movement is not attributable to a quality of the stimulus which is static, but an imposition of the perceiver. It has been argued

that phenomena such as movement are not perceptions (because blots are static, not moving) but mental constructions based on perceptual experiences (Blatt, 1992). In general, movement responses attempt to put greater specificity on form qualities. When such responses involve human activity, they often represent the individual's use of inner resources to modify perceptions of the external world. However, this interpretation changes if these efforts are combined with inaccurate form identification. If the forms identified differ greatly from those reported by others in a culture or society, then the individual is likely to make other people uncomfortable (without necessarily knowing why) by engaging in behaviors that depart from their expectations. Use of white space as figure-ground reversal is an analog to other figure-ground reversals such as an oppositional stance toward the world. Conventionality is interpreted according to the extent to which thought process is like others or different. The shading is a relatively subtle aspect of the Rorschach stimulus so that not all respondents report them as determinants (although they may perceive them). Such responses require a certain level of sensitivity or perceptiveness to verbalize.

The style of communication and nuances of language are important considerations (Smith, 1994). Therefore, characteristics of verbalizations such as redundancies or logical inconsistencies are noted. Even commonly reported responses can be expressed in unique ways. Percepts on the Rorschach are at times given without commitment or reflection, and the examinee might in such instances be reluctant to engage in the inquiry just as he or she was removed from the response process. These stylistic qualities of the response process provide a context for the interpretation of the variables in the structural summary.

Content given is also considered. For example, those who give a broad range of content have a greater variety of interests, and those who provide many human associations are exhibiting a broader interest in people. However, what is emphasized in coding responses is the manner in which the respondent uses the stimulus.

The Holtzman Inkblot Technique (HIT) (Holtzman, Thorpe, Swartz, & Heron, 1961) was designed to overcome what had been perceived as psychometric limitations of the Rorschach. Two parallel sets of inkblots were constructed, each set containing 45 inkblots plus two practice blots that are identical for both sets. The HIT was designed to differ from the Rorschach in several ways besides the number of inkblots (Holtzman, 1981). Two important

differences are that the respondent is instructed to give only one response per card rather than leaving it open and that a short and restricted inquiry is given after each response. This overcomes criticisms of the Rorschach arising from the variations in the style of inquiry and from the widely varying number of responses obtained which make it difficult to use norms. However, abandoning the principles basic to the Rorschach also has disadvantages. One cannot observe whether the respondent can see something different in an inkblot once it has been identified in a particular way (flexibility), and the number of responses *per se* is an important consideration. Sequential analysis of multiple responses to the same stimulus is not possible in the Holtzman format.

Clearly, in devising variations on techniques, there are trade-offs that need systematic attention. Given such acknowledgment, clinicians can make appropriate choices.

4.16.10.2 Interpretation

The process of interpreting the Rorschach has distinct phases. First, the examiner codes the responses and organizes them into patterns by calculating the variables in the structural summary. These variables are then compared with norms to designate deviations from expected patterns. A strategy for interpreting the protocol is selected based on constellations of responses that depart from the norms. The final step involves the synthesis of the norm referenced response patterns according to the psychological process associated with them. This step also involves analysis of the details of content and of the quality of the verbalized responses. The comprehensive system provides clear guidelines for each step of the interpretation process (Exner, 1993), but the examiner integrates information from various sources to refine conclusions and formulate recommendations.

The comprehensive system began with an almost exclusive emphasis on formal, structural characteristics emphasizing the perceptual-cognitive nature of the task to match blot areas with objects (Weiner, 1994). However, language and other associations are considered important (Rappaport, Gill, & Schafer, 1968) and have become increasingly incorporated into the interpretive procedure (Exner, 1993). Smith (1994) cautions against acceptance of the Rorschach as a test with one correct method of interpretation. An exclusive focus on a single approach fails to acknowledge the potential contribution of those with a different perspective. Content interpretation including recent

attempts at codings of object relations (Lerner, 1992; Stricker & Healy, 1990) is difficult to reconcile with the emphasis on structural variables. Historically, these two types of data have been viewed through different theoretical perspectives. An alternative suggested by Weiner (1994) is to focus on distinctions within personality rather than on classes of data such as form versus content.

4.16.11 DRAWING TECHNIQUES

Drawing tasks do not provide a stimulus but do exhibit the dual nature of projective devices as problem-solving tasks that reflect internal representations. Drawing a person or a family requires the translation of a three-dimensional memory image into a two-dimensional graphic representation within the constraints of the respondent's artistic ability. This conversion of three dimensions into two and compliance with instructions are the chief problem-solving task demands. Performance of the drawing is affected by the conceptualization of the object drawn, motor execution, attention to detail and spatial relationships, as well as planning and organizing the production. Like other projective tasks, the drawing can be evaluated according to its structure (proportion, elaboration, consistency of detail), content (what is drawn), and style (line pressure, size, placement). Machover (1949) suggested that structural or stylistic elements of size, placement, quality of line, positioning, symmetry, elaboration, or shading are more reliable than contents such as body parts and clothing. Performance standards for drawings can be set such as respect for inside and outside boundaries of the persons drawn and coherence and balance among details. A stylistic quality such as perfectionism and frequent erasures can be compared to the eventual quality of the product. The sheet of paper sets the boundaries or limits for the drawing. Very small or very large drawings suggest that the individual has difficulty setting boundaries in relation to the environment. However, other influences such as impaired psychomotor functioning (e.g., arthritis) in older clients (Kahana, 1978) must be acknowledged. Furthermore, processes such as lack of planning contribute to the use of available space. The individual may draw a head that is too large (lack of planning) so that the drawing either is out of proportion or cannot be completed on the page.

Drawings have been used to estimate intellectual functioning and neurological status as well as to assess personality. The draw-a-person Test (DAP) widely used as a projective

technique to assess personality (Machover, 1949) is also used to measure mental development (Goodenough, 1926). Its usefulness as an indicator of concept formation (intelligence) in the developing child was based on how closely the drawing approximated life-like proportions and details. Basically, the degree of realism was the criterion reflecting the child's conceptualization of the outward world through correct depiction of space and perspective. The Goodenough (1926) Draw-A-Man Test and the Harris (1963) revision of some criteria included items that appeared more frequently with increasing age. However, even as emphasizing age related trends, Goodenough (1926) noted the following individual variations in drawing that were not linked to development: (i) detail dominated but with few ideas; (ii) unique depictions that seem comprehensible only to the drawer; (iii) suggestive of a flight of ideas; and (iv) contradictory combination of primitive and mature characteristics.

Koppitz (1968) also differentiated qualities of children's drawings that were not a function of age. These aspects of drawings were stylistic: (i) qualitative aspects of integration, symmetry, and shading; (ii) presence of unexpected characteristics; and (iii) absence of expected characteristics beyond various ages. These qualities were viewed as manifestations of concept formation that were impacted by motivational and emotional aspects of the personality rather than chronological development. Drawing tasks are frequently given to adults to assess neurological status based on the observation that brain damage interferes with the integration of spatial, perceptual, and motor responses needed to execute drawings (Swindell, Holland, Fromm, & Greenhouse, 1988; Mendez, Ala, & Underwood, 1992).

Currently, the most frequently used projective drawing technique is the Draw-A-Person (DAP) as developed by Machover (1949) and expanded by others (Hammer, 1958; Handler, 1985; Koppitz, 1968, 1984; Urban, 1963). Also popular are the kinetic family drawing (KFD) technique (Burns, 1987; Burns & Kaufman, 1970, 1972); and House-Tree-Person (H-T-P) drawing task (Buck, 1948, 1987).

4.16.11.1 Interpretation

The dichotomy between person centered and product centered interpretation mentioned earlier is relevant here. Person centered approaches organize aspects of the production according to relevant psychological processes of the client. The interpreter looks for coherence in form, content, and style of the production with the

hypothesized psychological variables. The professional also notes consistency with drawing instructions, nature of verbalizations about the figures drawn, and the amount of effort expended. Global interpretations of interrelated parts are more valid than isolated interpretation of specific details. The sign approach to validating drawings has been criticized (Kahill, 1984; Roback, 1968; Swensen, 1968) because there are multiple possible interpretations for any quality, depending on the overall pattern. For example, size and detail of drawings can signal preoccupation, differential valuing, conflict (Fisher, 1986) and/or quality of planning and organization. Similarly, qualities such as shading or erasures are said to signal anxiety. However, the nature of the shading is important since this feature may enhance artistic quality and not reflect anxiety. It should also be noted that people differ widely in how they cope with anxiety. Some anxious individuals will race through the task without shading or erasing. If the focus is on the psychological processes of the individual, then the configuration of response elements or signs would be expected to cohere around that process.

It has been suggested that the clinician ask what a particular sign could mean rather than what it does mean (Handler, 1985). A critical review of the literature on the KFD technique (Handler & Habenicht, 1994) emphasized the need to study more holistic, integrative approaches to the KFD rather than the interpretation of a series of single signs. The authors argue for the importance of focusing research on the interpretive approach of the clinician using the KFD rather than on the technique itself. The use of drawings for multiple purposes (e.g., concept formation, neurological status, or personality) demonstrates the multidimensional aspects of projective tasks which permit simultaneous evaluation of responses from multiple perspectives to reveal various interrelated facets of functioning. Broad performance expectations pertaining to structural features of the product can be delineated. As with other performance measures of personality, individuals are free to vary their approach to meeting these general problem-solving expectations. One caveat is that artistic quality of the drawings appears to influence clinician's interpretation (Feher, VandeCreek, & Teglassi, 1983).

4.16.12 CASE ILLUSTRATION

On the surface, the three projective techniques reviewed seem very different. Yet, they reveal similar information when compared in terms of structural qualities. The overlap

between the TAT and Rorschach is illustrated by comparing formal aspects of Carl's TAT stories with conclusions drawn from the Rorschach. Carl, a 19-year-old man, had stopped attending school on a regular basis after the seventh grade and left school altogether during the 10th grade. At the time of this evaluation, Carl was incarcerated and attempting to continue his education in prison. He was doing poorly in his classes and was referred to determine if he would qualify for special education services. His WAIS-R IQ score was

low average. The intent here is not to develop a comprehensive clinical picture but to point out the overlaps between the two projective techniques. Table 2 shows the consistency of the TAT variables across cards and also displays the relevant Rorschach data.

Card 1. Little boy thinking, he's thinking how. I guess he's thinking a way how to work the violin. (Before?) He don't know how to work the violin. (TO?) Turns out he's still sitting there thinking.

Table 2 Corresponding aspects of Carl's TAT stories and Rorschach variables.

<i>TAT variables</i>	<i>Card</i>									
	1	2	3	4	5	6BM	8BM	12M	13MF	13B
Imprecise accounting of the stimulus		X	X	X		X	X		X	
Vague, concrete, or stereotypic story	X	X	X	X	X	X	X	X	X	X
Requires more than one query	X	X	X	X	X		X	X	X	X
Story ending is concretely tied to the stimulus	X	X	X	X	X	X	X	X	X	X
Relationship among characters are unclear, stereotypic, or unstated	NA	X	NA	X	NA	X	X	X	X	NA
Sense of helplessness (e.g., inaction, when it is warranted; lack of initiative or inertia)	X		X	X	X	X		X	X	X
Inner life such as intentions for actions or feelings is not sufficiently elaborated	X	X	X	X	X	X	X	X	X	X
Implausible sequence of events (e.g., cause-effect, timing, coherence of action with purposes)	X	X	X	X	X	X	X	X	X	X
Focus is dominated by the immediate circumstance or consideration	X	X	X	X	X	X	X	X	X	X
Insufficient integration of detail to comply with instructions	X	X	X	X	X	X	X	X	X	X

Rorschach variables

Resources and control

EB = 0:0; EA = 0; eb = 7:0; es = 7; *D = -2; *Adj. D = -2

Affect

FC:CF + C = 0:0; Pure C = 0; *Afr = .2143; S = 0; *Blends: R = 0:17; CP = 0

Interpersonal

* COP = 0; AG = 0; Food = 0; Isolate/R = .1764; * H:(H) + (Hd) + Hd = 0:0 *H + A:Hd + Ad = 13:0

Ideation

a:p = 4:3; Sum 6 = 0; M^a:M^p = 0:0; Lvl 2 = 0; 2 Ab + (Art + Ay) = 0; W Sum 6 = 0; M- = 0; M none = 0

Mediation

*P = 4 *X + % = .29; *F + % = .30; *X-% = .4117; S-% = 0; *Xu% = .29

Processing

Zf = 12; *Zd = -8.5; *W:D:Dd = 13:1:3; *W:M = 13:0 *DQ + = 3; *DQv = 4

* Deviates from normative expectation; R = 17; Lambda = 1.428; PSV = 3; Positive indices: SCZI; CDI.

The boy in the story can only keep sitting there thinking how to work the violin. He does not have sufficient resources to deal with the task or to find alternatives. Likewise, the narrator cannot go beyond the picture cues provided (concrete) to produce an ending or to describe purposes or deliberate actions. When prompted to tell what happened before, the response (didn't know how to work the violin) stayed within the moment without considering larger purposes. The fact that the narrator does not introduce other characters to garner support is consistent with the absence of human content in the Rorschach. The helplessness displayed by the character and the narrator is consistent with the positive Coping Deficit Index (CDI) on the Rorschach, suggesting insufficient resources to formulate responses to demands. The short and simplistic story is devoid of inferential or interpretive processes as consistent with vague and concrete information processing on the Rorschach (Lambda, Zd, DQ+, DQv, W:M).

Card 2. The man in the field working. Lady holding onto the stump because she looks like she's pregnant. Girl just come home from school. (T&F). Probably thinking about all the work they gotta do. And the man and the girl probably feeling the same pain, or sorta, for the lady that's pregnant. (Before?) The man started working. (TO?) Everything got done.

The story is closely anchored to the stimulus, yet the narrator does not give priority to the young woman in the foreground, nor does he indicate how the three people are related. Connections are vague and concrete such as the description of the lady as holding on to the stump because she's pregnant. Characters are differentiated only by external appearances, and they all feel the same. As in the previous story, inner life is not elaborated. Carl's Rorschach record contains no human movement, a variable that is associated with greater acceptance of inner thoughts and greater interpersonal awareness. None of the TAT stories suggests the availability of these resources.

Card 3. Little kid was probably tired so he or she fell out the chair. (T/F?) Maybe, I guess they think that's where they gonna sleep at. And probably feeling stupid because he's on the floor. (TO?) That's where they gonna be.

The character falls because he's tired and, despite feeling stupid, just stays there. Again, there's no purposeful action or inner resources but an inertia that's consistent with vague and insufficient processing of information. Feeling

stupid, yet doing nothing, is consistent with low self-worth shown on the Rorschach. Other Rorschach variables suggest that self-perceptions are naive and not guided by insight (no FD, no H) nor are these perceptions accompanied by dysphoric mood, anger, or irritation.

Card 4. Lady trying to convince her husband that she love him, but he ain't trying to hear it. (Before?) They was fussing, probably. (TO?) He's about ready to walk away.

The story shows no attempt to understand the inner life or concerns of others but indicates detachment from relationships. The explanation of the stimulus is incongruous—the man looks very angry, but his wife is declaring her love. When asked what happened before, the couple was described as fussing, but the nature of their disagreement is unstated. The picture shows the man turning away, and the narrator does not have the inner resource to provide a solution to the conflict that departs from the stimulus. The Rorschach also suggests that social relationships are superficial, distant, and guarded. The oversimplifying style is likely to lower sensitivity to the needs and interests of others, and this tendency is coupled with a lack of interest or detachment from relationships (no H; no COP or AG). The affect cluster suggests an approach to the environment that minimizes affective engagement. Lack of resources rather than affective provocation is at the heart of Carl's difficulties. His reluctance to process emotional stimuli (Afr) is consistent with his general style of oversimplified processing of information (low blends, high lambda, underincorporating approach).

Card 5. The lady must have heard someone in the house, so she ran and hide in the closet. (T/F?) She probably thinking that she gonna get hurt. (TO?) That no one was there.

The woman runs and hides because she feels vulnerable. Later, when she realizes there's no danger, she makes no attempt to discover the source of her concern (noise). This absence of initiative and curiosity is consistent with non-reflection, simplified processing, and detachment evident in the Rorschach as described previously.

Card 6BM. Looks like his mother told him some bad news. But the news didn't only hurt him; it hurt both of them. (TO?) That something did happen.

Again, the story offers vague, nonspecific descriptions such as bad news and little differ-

entiation between characters. Since the nature of the bad event is not understood, it is not possible to deal with it adaptively. The absence of specific story details is consistent with the other cards and with vague, simplified processing suggested by the Rorschach. The story ends with the character convincing himself that something did happen.

Card 8BM. Looks like they was trying to rob the lady, and she shot one of them. And the man's two buddies trying to get bullets out of him. (T/F?) Probably thinking that he ain't gonna make it. (TO?) That he made it.

The story presents an unrealistic sequence of events and poor integration of foreground and background components of the picture. There's no reasoning about intentions or consequences. Rather, the narrator gives simplistic associations to the stimulus without rule governed connections among likely sequences of events, between causes and effects, or between short and long term outcomes. This detachment from conventional thinking is consistent with poor reality testing on the Rorschach (poor form quality, low populars, incomplete processing of available information).

Card 12M. This is a lady? (It can be whatever it looks like to you) The old man is praying for the lady because he thinks that if he prays, she'll get better. (Before?) She was sick. (TO?) That she didn't make it.

As with the other stories in this protocol, the story does not depart from the immediate cues of the stimulus. Connections between sequences of events and among characters depicted are vague or nonexistent. For example, we do not know what is the relationship between the old man and the lady. The narrator is powerless to introduce alternatives that are not cued by the stimulus such as consulting a physician. The lack of integration and insufficient organization of details is consistent with Rorschach patterns suggesting that Carl formulates decisions without sufficient processing of information (e.g., high lambda, low developmental quality, low blends, and underincorporating style).

Card 13MF. The old man just got up for work, but his wife didn't have breakfast made for him. (T/F?) One of them feeling really tired. The other feeling left out. (TO?) Turns out that one of them sleep and one of them woke.

This story is also rather concrete and stereotypic. Each character has his and her roles to

play, but the narrator cannot develop the story beyond what is seen in the stimulus (reliance on immediate external circumstances).

Card 13B. Little boy ain't got no friends and wants something to drink bad. (Before?) Just sitting there. (T/F?) I guess he thinking why he's the only one there. (TO?) That he wasn't.

Just as in Card 5, the ending contradicts the initial premise of the character. Things are not the way they seem, but there's not enough initiative to investigate or deal with the discrepancy. Such lack of investment in processing information is compatible with the failure to process critical cues (underincorporation) on the Rorschach. Again, no purpose, interpersonal connection, or guiding principle is expressed.

The similarity of the conclusions derived from both techniques suggests that they are assessing common processes. Both methods suggest that resources and reality testing are not adequate to meet daily life demands. Furthermore, the focus on the immediate and the concrete, along with haphazard processing of information, hinder the development of schema that are sufficiently elaborated to provide rules that govern the synthesis of experience, the regulation of behavior, and the expression of affect. Without such inner guides to self-regulation, Carl exhibits tendencies toward impulsive, antisocial behavior and has limited capacity to delay gratification. Carl would benefit from a highly structured learning environment with highly structured tasks and frequent feedback and redirection.

4.16.13 VALIDATION ISSUES

Major criticism has been directed not only towards projective methods but also towards psychometric approaches to assessing personality which classify instances of experience, thought, or action into trait categories. The inference process, in general, has been maligned by the behaviorist movement. Compelling arguments have been made to exclude from consideration anything but overt behavior and objectively coded environmental variables (Skinner, 1953). Psychology has moved far from that position to a recognition of the importance of understanding the meanings that individuals assign to environmental events and to their own behaviors. Indeed, cognitive structures and processes are now considered to be at the heart of individual differences in experience, thought, and action, both adaptive and maladaptive (Cantor & Kihlstrom, 1987).

The process of inference is not restricted to the use of projective tests—it is what makes the psychologist a professional rather than a psychometrist. By giving priority to some scores or using qualitative information to increase understanding of the scores, the clinician engages in the process of making professional judgments (Groth-Marnat, 1990). These judgments cannot and should not be eliminated. It is important, however, for practitioners to reflect on the quality of their inferences and the usefulness of their decisions.

Evidence for validity of conclusions made on the basis of inferences drawn from projective methods needs to be convincing. Absolute standards that cut across all methods are essential and they must be precise enough to inform judgments about when the standard is or is not being met. However, in their application to projective techniques, general criteria should accommodate to the nature of the method. Validity and reliability are not established for a generic technique such as figure drawings, the TAT, or the Rorschach. Rather, the utility of each set of instructions and stimuli together with the scoring method is separately evaluated according to the accomplishment of its specific purpose. Ways of demonstrating validity and reliability for all techniques must be in tune with the logic and coherence of the measures and constructs under consideration.

4.16.13.1 Reliability

Establishing reliability for projective techniques and for questionnaire methods is fundamentally different. As Kelly (1958) pointed out: “When the subject is asked to guess what the examiner is thinking, we call it an objective test; when the examiner tries to guess what the subject is thinking, we call it a projective device” (p. 332). Reliability of items on a rating scale is essentially a matter of consistency in the respondent’s interpretation of the items. Reliability of inferences based on projective methods rests on a combination of the stimulus, the response, the method of interpretation, and the skill of the interpreter. Psychometric procedures to establish reliability based on forced-choice test items can be carried out without any clinical expertise because reliability pertains primarily to the test-takers responses. With projective measures, reliability, in part, is an attribute of the interpreter because the scoring units are inferences of the professional. Attempts to designate cookbook procedures so that the validation process can be carried out quickly and easily are out of tune with the nature of projective methods.

The following is a brief overview of how traditional psychometric indicators must be modified to establish reliability of projective techniques:

4.16.13.1.1 Scorer reliability

The accuracy of two people looking for the same information requires adequate training and clear guidelines. When raters are well-trained and scoring systems are well documented, such reliabilities tend to be high. A related aspect of scorer reliability, particularly relevant to clinical inference is the consistency with which one rater codes the same protocol over time (Karon, 1981).

4.16.13.1.2 Decisional reliability

The consistency of decisions drawn from a protocol can be assessed when specific units are not the focus. For example, Shneidman (1951) showed that 16 clinicians using their own methods came to similar conclusions. Other influences on the reliability of decisions may relate to the number of performance samples needed. For example, in the measurement of particular motives with TAT cards, it is important to know if the specific motive emerges reliably in every card or only some. To assure adequate reliability, it has been recommended that at least six stories be obtained from each respondent (Lundy, 1985; Smith, 1992b).

Decisions must be based on clear conceptualizations. Therefore, decisional reliability must account for the theoretical appropriateness of the match between predictor and criteria. Test responses can be expected to correlate with criteria only if their meaning is functionally similar. Thus, adequate functioning in a structured situation can occur despite disorganized responses on a projective test.

4.16.13.1.3 Test–retest reliability

One important factor in demonstrating the reliability of the measure upon retesting is whether one is looking for similarity of content or of the inference (Karon, 1981). The reliability of the specific content is less relevant than consistency in the meaning of the response.

4.16.13.1.4 Internal consistency

Internal consistency of thematic content across cards would be an inappropriate measure of reliability because cards are designed to elicit different themes (Lundy, 1985). However, despite variability in specific content, the responses may yield similar inferences. More stylistic units representing psychological processes such as the accuracy with which the

content captures the “gist” of the scene presented or linkages between causes and effects may be generalizable across different stimuli. Internal consistency based on the number of words per story ($\alpha = 0.96$) was much higher than internal consistency of need for achievement (Atkinson, Bongort, & Price, 1977). Likewise, Rorschach cards present stimuli with important differences and, rather than estimating internal consistency, inferences are made on the basis of the entire coded protocol and not card by card. Responses to the white spaces in the blot have different meaning for each card, and not all white space responses represent figure ground reversal. Therefore, adding all space responses is a rough estimate of the trait in question.

Internal consistency expected within a measure should relate to the nature of the task and to the consistency inherent in the construct under consideration.

4.16.13.2 Construct Validation

Messick (1989) defines construct validity as “an integration of any evidence that bears on the interpretation or meaning of test scores” (p. 17). Because traditional indices of content or criterion validity contribute to the meaning of test scores, they too pertain to construct validity. Thus, construct validity subsumes all other forms of validity evidence.

This emphasis on construct validity as the overriding focus in test validation represents a shift from prediction to explanation as the fundamental focus of validation efforts. Construct validation emphasizes the development of models explaining processes underlying performance on various tests and their relationships to other phenomena. Accordingly, correlations between test scores and criterion measures contribute to the construct validity of both predictor and criterion. In Messick’s words, “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (p. 13).

4.16.13.3 Multitrait, Multimethod Validation

Although the multitrait, multimethod approach is a standard technique for construct validation (Campbell, 1960; Campbell & Fiske, 1959), this technique is vulnerable to problems with the definition and measurement of constructs. This procedure seeks to establish higher correlations across diverse measures of the same trait (convergent evidence) and lower correlations among similar measures of different traits (discriminant evidence) to show that a construct

is distinct from other constructs and that it is not uniquely tied to a particular measurement method. Yet, the possibility that the constructs may indeed be tied to the measurement tool must be considered. McClelland and colleagues showed that attempts to correlate self-report and projective measures of achievement motivation to validate either one are misguided. Likewise, Horowitz (1991) notes that schema, although important to measure, are not directly known by the subject and cannot be confirmed through self-report but by a consensus of two or more independent observers. Theoretical understanding of the construct must always be central to the choice of validation efforts.

It is important to acknowledge the potential for corroboration of constructs with various assessment measures. Yet, different measures of the same construct (e.g., ratings by actor versus observer) may be assessing qualitatively different aspects of that construct. Meaningful differences in the constructs are demonstrated by distinct patterns of relationships with various criterion measures (see criterion validation below).

Even highly similar measures (e.g., thematic approaches) used to assess identical constructs but with different procedures are not necessarily addressing the same qualities. For example, Arnold (1962) developed criteria for scoring TAT stories for achievement motivation by comparing groups of individuals known to differ in their job success. In contrast, the McClelland–Atkinson tradition (McClelland & Koestner, 1992) for developing criteria for scoring achievement motivation from stories told to picture stimuli was based on the comparison of groups given different instructional sets to arouse the achievement motive. Each of these two approaches to contrasting groups may be appropriate for given purposes with specific populations but clearly involve distinct conceptualizations of the achievement motive. One takes the position that the motive is present and, when aroused, energizes and directs behavior towards a particular class of goals or incentives. The other focuses on the achievement motive as being a relatively stable disposition reflected by complex patterns of cognitive–emotional processes that guide perceptions and behavior. If scoring units are empirically derived through contrasting groups, then the construct is defined by the nature of the groups and procedures used.

4.16.13.4 Criterion Validation

Criterion related evidence also belongs under the rubric of construct validity (Messick, 1989). Patterns of correlations with other variables

contribute to the defining attributes of each construct. The same set of external correlates do not apply to achievement motivation measured through projective and self-report methods (Spangler, 1992). Projective measures of achievement motivation correlate with spontaneous effort sustained over time, whereas self-reports correlate with activities that are cued by the situation. Such variation in the match between the predictor and the target clarify the dimensions of the achievement motive assessed by each type of measure. This understanding of constructs in terms of the life conditions to which they generalize constitutes the simultaneous validation of the predictor and the criteria. Relationships between predictors and criteria derived from a construct must be tested for expected patterns of discriminant and convergent evidence. In doing so, both situational specificity and generalizability of constructs and their measures are addressed at the same time. Generalizability of the predictor–criterion relationships across different population groups, settings, task domains, ages, or gender must be empirically established.

4.16.13.5 Part–Whole Relationships

Personality is a functional whole comprised of patterns of interrelated components. Therefore, predictions from single elements of personality to more encompassing units, such as general adjustment, proceed from part to whole and can only account for a portion of the variance. This part–whole prediction is illustrated in Figure 1 showing the influence of temperament on adjustment as mediated by goodness-of-fit. Any one temperamental attribute would account for a small proportion of the variance in the fit between the person and various situational contexts or task demands. A configuration of such traits would improve the prediction. However, the actual goodness-of-fit would be a better predictor of adjustment because this broader construct subsumes the configuration of traits (even those not specifically addressed in the prediction), the environmental demands and supports, as well as coping mechanisms.

This understanding of the part–whole relationship is essential for the validation of projective techniques. Projective techniques look at individual differences at a more global level than other measures. The products reflect schema or inner structures that develop through the individual's synthesis of life experiences. These involve the interplay of all of the person's characteristics and reciprocal transactions with the environment. Therefore, part–whole patterns exist between various trait constructs and response parameters within the test. Part–whole predictions can relate specific neuropsychological or temperamental processes to individual differences in responses to projective tasks (Bassan-Diamond, Teglassi, & Schmitt, 1995). Part–whole relationships depend on conceptualizations about how the various dimensions of individual differences come together in the development of broader units of personality such as the influence of temperament on conscience development (Kochanska, 1993). Given the expected role of multiple variables, it is apparent that expectations about the size of correlations need to be in line with understanding of the phenomena.

The part–whole issue also applies to the relationships among responses within a specific projective technique. The functional significance of each part hinges on its interplay with other parts in relation to the constructs being considered. Various dimensions of the product such as form and content variables are cohesively related and need to be systematically coordinated into higher-order constructs. Psychometric and theory-based approaches need to be integrated to build useful frameworks for clinical use. Specific response parameters can be abstracted from the whole and pieced together to find conceptually meaningful patterns.

4.16.13.6 Convergence of Psychometric and Conceptual Treatment of Data

Normative, nomothetic, and case study approaches to data collection are interrelated. Conceptual linkages among these approaches are essential to the interpretation and validation of projective techniques, as described next.

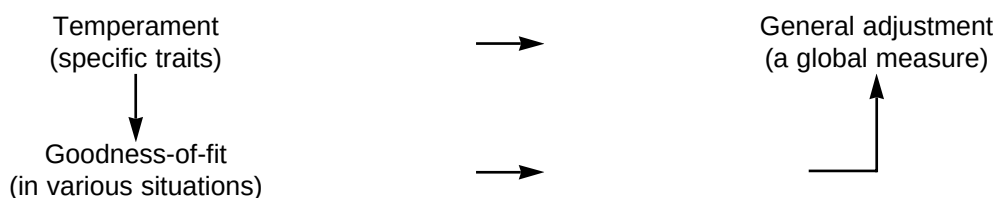


Figure 1 Illustration of part–whole prediction.

4.16.13.6.1 Normative

Before initiating the arduous process of collecting norms for projective methods, the units of analysis must be conceptually meaningful, theoretically relevant, and clinically useful. Furthermore, criteria for coding must be provided in sufficient detail to assure rater reliability. The Comprehensive System for the Rorschach has shown that the establishment of norms for meaningful units of analysis is possible, and that larger conceptual units can be designated through the use of multiple cut-offs to establish interpretively useful patterns of interrelated response parameters.

4.16.13.6.2 Nomothetic

It is important to base norms on units of interpretation that represent psychological processes that can serve as explanatory constructs. The nomothetic approach attempts to promote conceptual understanding through the study of general principles and functional relationships among variables. Such a conceptual approach to the Rorschach (Blatt & Berman, 1984; Weiner, 1977) has advocated organizing discrete response parameters into theoretically relevant clusters that contribute to a well defined construct. A strict empiricist would be satisfied with the prediction of patterns of behavior from various configurations of test responses (e.g., of Rorschach variables). However, establishing such empirical relationships is only a starting point for building conceptual frameworks for understanding and explaining behavior. The meaning of test response patterns emerges from a network of relationships with other theoretically relevant response patterns and theoretically appropriate external criteria.

The collection of normative data for projective techniques is embedded in the construct validation process. The constructs and units of inference for projective measures need to be defined *a priori* in the same way that test items in questionnaires are selected to represent constructs. An example of this approach is the designation of units of inference for TAT stories on the basis of constructs derived from research and theory on empathy (Locraft & Teglassi, 1997). Scores based on these units subsequently differentiated groups of children designated as high, medium, and low on empathy on the basis of teacher ratings. Such units require cross validation on numerous populations to assure wide applicability prior to attempting large scale normative studies. Unless the unit of inference is conceptually clear and represents meaningful psychological processes, normative data or group comparisons are not particularly useful.

4.16.13.6.3 Case study

The case study is most relevant to the practicing clinician because decisions for clinical purposes relate to understandings about one person. The adequacy of these decisions can be validated on a case by case basis. Although case study methods focus on the unique patterns of variation in responses of a single individual, the pattern of variables under consideration can be referenced to norms and to theoretical constructs. With projective methods, the respondent is allowed free expression within a specified context (stimuli and directions), and effective use of these techniques require a conceptual framework at three levels. First is an understanding of the psychological meaning of the response patterns within the projective measure. Second is an understanding of how the projective measure fits with other information in a comprehensive evaluation. Third is a conceptualization of how the patterns relate to competencies needed in the relevant life situations. Therefore, to conduct adequate case studies, the examiner needs a framework to understand how various types of data fit together within and across measures and how the emerging patterns apply to the various demands of the individual's environment. The attempt to integrate psychometric (empirical) and nomothetic (theoretical) data into the case study is typified in the attempt to synthesize structural and content features of Rorschach responses (Erdberg, 1993). The problem is that these efforts have applied different conceptual frameworks to different aspects of the data. A true integration of the empirical and theoretical perspectives requires a focus on personality constructs and the establishment of coherent patterns of data from various sources around the constructs.

The term "conceptual validity" has been applied to the process of psychological assessment (Maloney & Ward, 1976). Whereas construct validity focuses on confirming expected patterns of relationships across individuals, conceptual validity focuses on observing cohesive patterns within an individual. These patterns of expected relationships among observations constitute a working model of the individual being evaluated. According to Maloney and Ward, establishing such a model is a prerequisite for answering the referral question. When information from various sources is understood in terms of the constructs that explain an individual's difficulties and point to appropriate decisions, the assessment has conceptual validity.

4.16.14 FUTURE DIRECTIONS

The anticipation of future directions for projective techniques emerges from the clues gleaned from the scientific literature pertaining to the development and assessment of personality. However, the prognosticator's wishful thinking also influences this process. The intent of the prognosticator is to show how unfolding trends or even new spins on old ideas point toward desired alternatives for moving forward. Future possibilities for projective techniques are drawn from three conceptually distinct perspectives: (i) converging constructs from the various subfields of psychology that scaffold the use of projective techniques; (ii) refinements in the conceptualization of personality that point to a multidimensional view and to a unique role for projective techniques; and (iii) improved understanding and use of the specific projective tools.

Methods and constructs in other subfields of psychology provide a hospitable *zeitgeist* for projective techniques and bode well for their future development and utility. Converging evidence appears to validate the basic tenets of the projective hypothesis. Concepts from psychodynamic formulations such as transference have been redefined in terms of contemporary psychology, demonstrating that such phenomena are not unique to one theory but can be understood in several ways (Singer & Singer, 1994). The examination of clinically relevant phenomena in light of conceptualizations of memories, schema, and scripts from various subdisciplines support the work in each subfield (e.g., Horowitz, 1991; Stein & Young, 1992). Cognitive theories of perception, memory, and learning are increasingly emphasizing the unconscious or implicit social attitudes (Uleman & Bargh, 1989). Research on construct availability and accessibility (Higgins, 1990) as well as script or schema theory also recognizes the influence of cognitive processes that occur outside of awareness. Knowledge from other areas of psychology will increasingly inform the designation of interpretive units for measuring personality with projective techniques. Recent interest in narrative methods to assess schema and social cognitions is applicable to thematic apperceptive techniques (Cramer, 1996). Stories elicited through picture stimuli, like other narrative approaches, reflect the human orientation to perceive the world as stories or myths grounded in culture and personal experience.

Conceptions of psychopathology within psychoanalytic theory are placing increasing emphasis on functions of inner psychic structures and subjective meaning systems (Atwood & Stolorow, 1984). In this context, adjustment problems arise because the structures are

insufficiently developed to permit adaptive coping rather than from conflict among the structures. In an analogous manner, clinical syndromes have been related to problems in the development of person schema (Horowitz, 1991). If psychopathology is a reflection of impairment in the formation of psychic structures, then it is reasonable for assessment to be geared to the evaluation of these structures (e.g., object relations, self-system, along with associated processes such as reality testing).

Levels of impairment based on the complexity and organization of inner structures will be identified along with a description of symptomatic behaviors. Projective techniques will figure prominently in the assessment of psychological variables for the purpose of intervention planning. The continued refinement of integrative therapies (Leve, 1995; Norcross, 1986) will increase the usefulness of identifying relevant processes to be targeted and matched with optimal intervention strategies.

Understanding personality according to different styles and levels of organization of experience as they relate to functioning will rest on increasing emphasis on part whole conceptualizations. For example, with a focus on inner structures, aspects of personality such as emotions will be increasingly recognized as organizing processes that shape adaptation and problem solving (Greenberg, Rice, & Elliott, 1993). Other personality processes, such as distractibility or inattention, will also be understood in relation to how they shape the development of inner psychic structures. Future research will focus on establishing linkages between various levels of personality constructs such as the role of various traits in the development of inner strivings and the organization of meaning structures. Research will also focus on clarifying relationships between discrete measures of neuropsychological processes, such as planning, organizing, retrieval from memory, or continuous attention, and the manifestation of these processes in the performance of more complex tasks analogous to real-life situations. Inherent in the part whole conceptualization of personality is the view of development as being propelled simultaneously by biological, psychological, and social forces. Within the psychological realm, perceptions and behaviors will be increasingly understood in terms of the interplay of experience with relevant affective and cognitive processes.

A multilevel understanding of personality will permit clearer conceptualizations of the utility of different types of techniques to assess different facets of personality. One example is the distinction between measures of self-attributed and implicit motives to achieve (McClelland

et al., 1989). Story-telling measures of achievement motivation (implicit) were more effective in the prediction of long term outcomes such as career success, whereas self-reports (self-attributed) were better in predicting immediate choices (Spangler, 1992). The two types of measures of achievement motivation relate to different criteria, develop through different pathways, and can be understood in terms of different levels of personality.

In training programs, the professional courses are the places where the core psychological knowledge bases are integrated. Expertise in projective techniques includes the linking of the psychological processes assessed with appropriate strategies for therapeutic interventions. The complexity of interpreting projective instruments and of understanding their implications for guiding interventions requires no less than the systematic and flexible application of prior knowledge. Therefore, projective assessment can serve as a centerpiece for integrating concepts from core areas such as physiological, affective and cognitive bases of behavior. Training should emphasize the gradual development of the professional's schema to guide practice rather than the acquisition of knowledge applied in rote fashion. Trainees' schema should be sufficiently broad and flexible not only to incorporate information currently available but to accommodate continuous learning throughout the professional career. Measures are simply tools. Rather than emphasizing the assessment technique, training programs would do well to focus on the development of integrative frameworks that include how various sources of information fit together into meaningful patterns.

Increased acceptance of qualitative measures as having scientific merit has led to more frequent use of open-ended methods in the study of personality such as the study of early memories (Bruhn, 1992), thought sampling (Rubin, 1986), examination of the life story (McAdams, 1990), self-defining memories (Moffit & Singer, 1994), and analysis of therapy transcripts (Luborsky & Crits-Christoph, 1990). The psychometric challenges of projective methods are shared among all open-ended techniques. The effort to master these difficulties will require, above all, conceptual clarity and an emphasis on construct validation. These endeavors will promote shared conceptualizations and methods across disciplines, appreciation of the progress already made with projective techniques, and the spurring of further developments.

Frameworks for interpretation of projective techniques not only elucidate the inner structure of personality but reveal problem-solving

strategies in relation to the task demands. Therefore, it may be fruitful to view these techniques, in part, as *performance measures of personality* that can be compared to other tests in an assessment battery in terms of what the task requires. The most significant difference is that performance tasks of personality maximize the use of spontaneous strategies to organize perceptions and responses. Variation in response patterns across different types of tasks permits distinctions between knowledge structures that are inert and require external prompts from those that are meaningfully organized and spontaneously accessible. Furthermore, the configuration of responses within and across diverse tasks foster explanations of situational variability and consistency in performance and behavior. For instance, social situations require the individual to organize perceptions, size up intentions, and anticipate reactions of others. These requirements are similar to the demands of projective tests to organize perceptions of the stimuli and coordinate the responses with the instructions.

The relationship between content and formal elements in the units of inference drawn from projective techniques still needs to be clarified. The most fruitful approach to integrating the analysis of form and content is through the identification of their linkages with important psychological processes. Formal structural analysis should receive greater emphasis in the apperceptive methods. Generalized properties of content and features of narrative structure are more durable indices of psychological processes than isolated content. Furthermore, the analysis of the subtext or underlying structure of narratives can yield significant information about the individual's schematic organization of experience. The struggle to apply valid and reliable methods to assessment will continue. The inevitable link between theory and method necessitates the simultaneous effort to refine both. Conceptualizations developed in other subfields of psychology and lessons learned from qualitative research methods will be applied to improve the reliability, validity, and clinical utility of projective measures.

4.16.15 SUMMARY

The chapter is summarized by presenting a model for understanding performance measures of personality shown in Figure 2. The interaction of the person and environment (boxes 1 and 2) as a fundamental unit of study has wide acceptance and applicability. However, the objective features of the environment and the subjective

world are not the same. Lewin (1935) refers to the psychological environment as the inner experience of external reality. Persons participate in and influence their external environments as well as being shaped by them. The manner in which the individual encodes and stores encounters with the external world drives the construction of the inner psychological world (box 3). Therefore, one can only conceptually separate the person from the environment—they are embedded in one another.

Schema are memory structures that develop through the individual's constant and reciprocal transactions with the environment. The development of schema is influenced by (all factors said to influence personality) various interactive determinants including constitutional factors of temperament (see reviews by Emde, 1989; Plomin, 1986) as well as meaning structures transmitted by family and culture in conjunction with the individual's experiences throughout life. These schema are important to assess because they: (i) are mental sets that guide attentional processes and serve as filters for interpreting information; (ii) guide actions in situations that require complex resources; and (iii) shape the storage of new information in memory.

Certain conditions such as attention-deficit hyperactivity disorder are diagnosed through a careful examination of previous history as well as scrutiny of current behavioral patterns. However, these attentional processes shape the individual's prior and ongoing synthesis

of experience through their impact on the information entering awareness, the feedback received from others and on the development of internalized schema. Therefore, the part-whole relationships between specific symptoms or psychological processes and larger units of the personality need to be understood. Projective testing relies on an appreciation that overt behavior is linked to inner meaning. Consequently, inner life is a mechanism for perceiving the outer world and for regulating behavior to adapt to these perceptions. The projective techniques (e.g., TAT and Rorschach) provide information about how prior experience has been organized and how inner resources (schema, inner structures) are applied to relatively unstructured situations. Responses to projective tasks are understood as cohesive products reflecting all of the processes involved in the synthesis of experience. Specific psychological processes such as attention or emotion can be studied as precursors and sequelae of the development of inner structures.

Projective techniques require a person to demonstrate the qualities of organization and strategic planning by performing a problem-solving task rather than by telling about the self in an interview or by responding to a questionnaire. Therefore, the term *performance measure of personality* aptly describes projective methods. These performance measures are less structured than the other tasks in the typical assessment battery and provide unique information. The manner in which the problem is solved

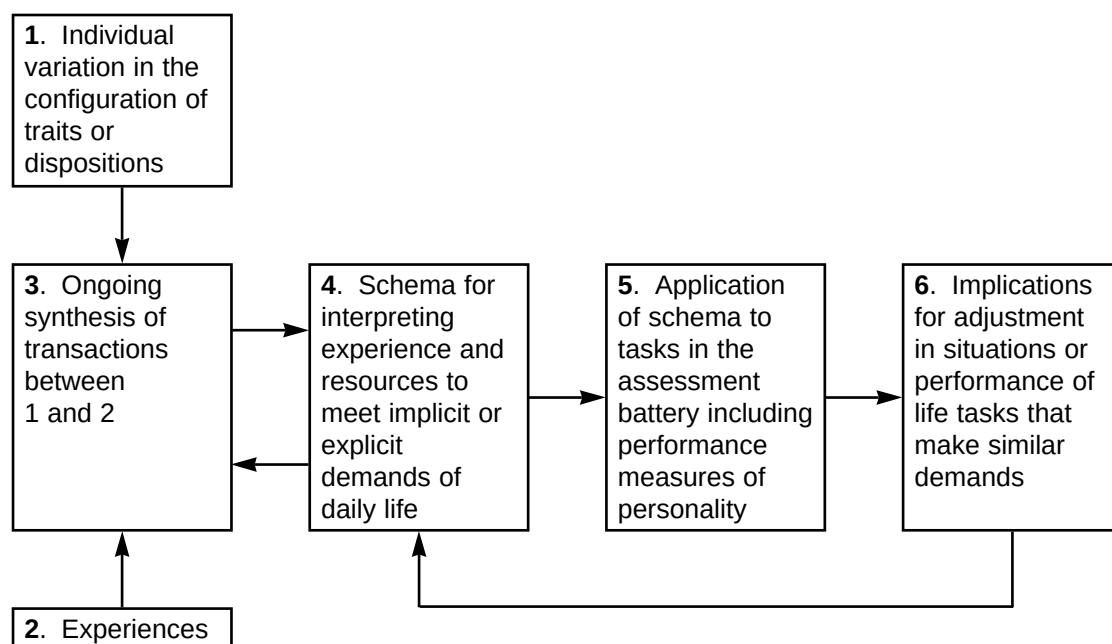


Figure 2 Understanding performance measures of personality.

or the task is accomplished reveals the individual's organizational strategies and resources to deal with similarly unstructured situations. In addition, the nature and organization of knowledge structures reveal the frameworks or convictions that guide responses in ambiguous situations. Furthermore, inferences drawn from performance measures and self-reports pertain to different facets of personality and predict different criteria (McClelland et al., 1989).

As Figure 2 shows (box 4), the individual's ongoing synthesis of experience leads to the development of knowledge structures. Various performance measures in an assessment battery can be examined in terms of what the task requires and what inner structures or resources are brought to bear on the performance (box 5). Actual adjustment in a given situation (box 6) depends not only on the person's resources but also hinges on environmental expectations and supports.

The person is assessed within a particular social context of the testing situation including expectations about testing, anticipation of the potential consequences of testing, and the relationship with the examiner. Therefore, these variables are considered along with the task demands. The job of the examiner is to organize multiple sources of data including reports of self and others, various performance measures, relevant past history, and current circumstances into a framework for understanding the referral issues. All of the information is sifted through an understanding of family and cultural background as pertains to the individual being assessed. The understanding is informed not only by theories of personality, human development, and psychopathology, but also by a conceptualization of the various tasks in the battery and how they relate to performance in life situations. Findings are not a listing of isolated facts but conclusions based on cohesive patterns that provide an understanding of the problem and point to appropriate action.

4.16.16 REFERENCES

- Abelson, R. P. (1976). Script processing in attitude formation and decision making. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, NJ: Erlbaum.
- Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist*, 36, 715-729.
- Acklin, M. W. (1994). Some contributions of cognitive science to the Rorschach Test. In I. B. Weiner (Ed.), *Rorschachiana* (pp. 129-145). Seattle, WA: Hogrefe & Huber.
- Alper, T. G., & Greenberger, E. (1967). Relationship of picture structure to achievement motivation of college women. *Journal of Personality and Social Psychology*, 7, 362-371.
- Anastasi, A. (1976). *Psychological testing*. New York: Macmillan.
- Anderson, J. R. (1990). *Cognitive psychology and its implications*. New York: Freeman.
- Archer, R. P., Maruish, M., Imhof, E. A., & Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. *Professional Psychology: Research and Practice*, 22, 247-252.
- Arnold, M. B. (1962). *Story sequence analysis: A new method of measuring motivation and predicting achievement*. New York: Columbia University Press.
- Atkinson, J. W. (1992). Motivational determinants of thematic apperception. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 21-48). New York: Cambridge University Press.
- Atkinson, J. W., Bongort, K., & Price, L. H. (1977). Explorations using computer simulation to comprehend TAT measurement of motivation. *Motivation and Emotion*, 1, 1-27.
- Atwood, G., & Stolorow, R. (1984). *Structures of subjectivity*. Hillsdale, NJ: Analytic Press.
- Auld, F., Jr. (1954). Contributions of behavior theory to projective testing. *Journal of Projective Techniques*, 18, 421-426.
- Bargh, J. A. (1989). Conditioned automaticity: Varieties of automatic influence on social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended thought* (pp. 3-51). New York: Guilford Press.
- Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 1-40). Hillsdale, NJ: Erlbaum.
- Barkley, R. A. (1990). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment*. New York: Guilford Press.
- Bassan-Diamond, L. E., Teglassi, H., & Schmitt, P. (1995). Temperament and a story-telling measure of self-regulation. *Journal of Research in Personality*, 29, 109-120.
- Baumeister, R. F., Wotman, S. R., & Stillwell, A. M. (1993). Unrequited love: On heartbreak, anger, guilt, scriptness, and humiliation. *Journal of Personality and Social Psychology*, 64, 377-394.
- Beck, S. J. (1981). Reality Rorschach and perceptual theory. In A. I. Rabin (Ed.), *Assessment with projective techniques* (pp. 23-46). New York: Springer.
- Bellak, L. (1975). *The T.A.T., C.A.T., and S.A.T. in clinical use*. New York: Grune & Stratton.
- Bellak, L. (1993). *The T.A.T., C.A.T., and S.A.T. in clinical use* (5th ed.). Needham Heights, MA: Allyn & Bacon.
- Birney, R. C. (1958). Thematic content and the cue characteristics of pictures. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society* (pp. 630-643). New York: Van Nostrand.
- Blatt, S. J. (1991). A cognitive morphology of psychopathology. *Journal of Nervous and Mental Disease*, 179, 449-458.
- Blatt, S. J. (1992). The Rorschach: A test of perception on an evaluation of representation. In E. I. Megargee & C. D. Spielberger (Eds.), *Personality assessment in America: A retrospective on the occasion of the fiftieth anniversary of the Society for Personality Assessment* (pp. 160-169). Hillsdale, NJ: Erlbaum. (Reprinted from 1990 *Journal of Personality Assessment*, 55, 394-416).
- Blatt, S. J., & Berman, W. H., Jr. (1984). A methodology for the use of the Rorschach in clinical research. *Journal of Personality Assessment*, 48, 226-239.
- Blatt, S. J., Ford, R. Q., Berman, W., et al. (1988). The assessment of therapeutic change in schizophrenic and borderline young adults. *Psychoanalytic Psychology*, 5, 127-158.
- Blatt, S. J., & Lerner, H. (1983). The psychological

- assessment of object representations. *Journal of Personality Assessment*, 47, 7-28.
- Blatt, S. J., & Wild, C. M. (1976). *Schizophrenia: A developmental analysis*. New York: Academic Press.
- Blatt, S. J., Allison, J., & Feirstein, A. (1969). The capacity to cope with cognitive complexity. *Journal of Personality*, 37, 269-288.
- Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be taught. In S. Voshiadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 470-497). Cambridge, UK: Cambridge University Press.
- Bruhn, A. R. (1992). The early memories procedure: A projective test of autobiographical memory, Part 2. *Journal of Personality Assessment*, 58, 326-346.
- Bruner, J. S. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press.
- Buck, J. N. (1948). The H-T-P technique: A quantitative and qualitative scoring manual. *Clinical Psychological Monographs*, 5, 1-120.
- Buck, J. N. (1987). *The House-Tree-Person technique: Revised manual*. Los Angeles: Western Psychological Services.
- Burns, R. C. (1987). *Kinetic-House-Tree-Person drawings (KHTP)*. New York: Brunner/Mazel.
- Burns, R., & Kaufman, S. (1970). *Kinetic Family Drawings (K-F-D): An introduction to understanding children through kinetic drawings*. New York: Brunnel/Mazel.
- Burns, R. C., & Kaufman, H. S. (1972). *Actions, styles, and symbols in Kinetic Family Drawings: An interpretive manual*. New York: Brunner/Mazel.
- Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. In J. T. Spence, J. M. Darley, & D. J. Foss (Eds.), *Annual Review of Psychology*, 47, 87-111.
- Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait or discriminant validity. *American Psychologist*, 15, 546-553.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Campus, N. (1976). A measure of needs to assess the stimulus characteristics of TAT cards. *Journal of Personality Assessment*, 40, 248-258.
- Cantor, N., & Kihstrom, J. F. (1987). *Personality and social intelligence*. Englewood Cliffs, NJ: Prentice-Hall.
- Carlson, L., & Carlson, R. (1984). Affect and psychological magnification. Deviations from Tomkins' script theory. *Journal of Personality*, 52, 36-45.
- Chi, M., Glaser, R., & Farr, M. (1987). *Nature of expertise*. Hillsdale, NJ: Erlbaum.
- Clore, G. L., Schwarz, N., & Conway, M. (1994). Affective causes and consequences of social information processing. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 323-417). Hillsdale, NJ: Erlbaum.
- Cooper, A. (1981). A basic TAT set for adolescent males. *Journal of Clinical Psychology*, 37, 411-414.
- Cramer, P. (1996). *Storytelling narrative and the Thematic Apperception Test*. New York: Guilford Press.
- Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social-information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, 115, 74-101.
- Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.
- Demorest, A. P., & Alexander, I. E. (1992). Affective scripts as organizers of personal experience. *Journal of Personality*, 60, 645-663.
- Dodge, K. A., & Feldman, E. (1990). Issues in social cognition and sociometric status. In S. R. Asher & J. D. Coie (Eds.), *Peer rejection in childhood* (pp. 119-155). New York: Cambridge University Press.
- Durand, V. M., Blanchard, E. B., & Mindell, J. A. (1988). Training in projective testing: Survey of clinical training directors and internship directors. *Professional Psychology: Research and Practice*, 19, 236-238.
- Emde, R. N. (1989). The infant's relationship experience: Developmental and affective aspects. In A. J. Sameroff & R. N. Emde (Eds.), *Relationship disturbances in early childhood: A developmental approach*. New York: Basic Books.
- Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist*, 49, 709-724.
- Erdberg, P. (1993). The U.S. Rorschach scene: Integration and elaboration. In I. B. Weiner (Ed.), *Rorschachiana XIX: Yearbook of the International Rorschach Society* (pp. 139-151). Seattle, WA: Hografe & Huber.
- Exner, J. E. (1989). Searching for projection in the Rorschach. *Journal of Personality Assessment*, 53, 520-536.
- Exner, J. E. (1993). *The Rorschach: A comprehensive system: Vol. 1. Basic processes*. New York: Wiley.
- Exner, J. E., & Weiner, I. B. (1995). *The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents*. New York: Wiley.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50, 229-238.
- Feher, E., VandeCreek, L., & Teglassi, H. (1983). The problem of art quality in the use of the Human Figure Drawing Test. *Journal of Clinical Psychology*, 39, 268-275.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.
- Fisher, S. (1986). *Development and structure of the body image*. Hillsdale, NJ: Erlbaum.
- Fiske, A. P., Haslam, N., & Fiske, S. T. (1991). Confusing one person with another: What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology*, 60, 656-674.
- Fiske, S. T., & Taylor, S. (1991). *Social cognition* (2nd ed.). New York: McGraw-Hill.
- Frank, L. D. (1939). Projective methods for the study of personality. *Journal of Psychology*, 8, 389-413.
- Frank, L. D. (1948). *Projective Methods*. Springfield, IL: Thomas.
- Fromkin, H. L., & Streufert, S. (1976). Laboratory experiments. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 415-465). Chicago: Rand McNally.
- Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. New York: World Books.
- Greenberg, L. S., Rice, L. N., & Elliott, R. (1993). *Facilitating emotional change: The moment by moment process*. New York: Guilford Press.
- Groth-Marnat, G. (1990). *Handbook of psychological assessment* (2nd ed.). New York: Wiley.
- Hammer, E. F. (1958). *The clinical application of projective drawings*. Springfield, IL: Charles C. Thomas.
- Handler, L. (1985). The clinical use of the Draw-A-Person Test (DAP). In C. S. Newmark (Ed.), *Major psychological assessment instruments*. Newton, MA: Allyn & Bacon.
- Handler, L., & Habernicht, D. (1994). The Kinetic Family Drawing Technique: A review of the literature. *Journal of Personality Assessment*, 62, 440-464.
- Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity*. New York: Harcourt Brace Jovanovich.
- Hartman, A. A. (1970). A basic TAT set. *Journal of Projective Techniques*, 34, 391-396.
- Haynes, J. P., & Peltier, J. (1985). Patterns of practice with TAT in juvenile forensic settings. *Journal of Personality Assessment*, 49, 26-29.

- Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.
- Henry, W. E. (1956). *The analysis of fantasy: The Thematic Apperception Test in the study of personality*. New York: Wiley.
- Higgins, E. T. (1990). Personality, social psychology, and person-situation relations: Standards and knowledge activism as a common language. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research*, (pp. 301–338). New York: Guilford Press.
- Hogan, R. (1987). Personality psychology: Back to basics. In J. Aronoff, A. J. Rabin, & R. A. Zucker (Eds.), *The emergence of personality* (pp. 79–105). New York: Springer.
- Holt, R. R. (1958). Formal aspects of the TAT—A neglected resource. *Journal of Projective Techniques*, 22, 163–172.
- Holt, R. R. (1961). The nature of TAT stories as cognitive products: A psychoanalytic approach. In J. Kagan & G. Lesser (Eds.), *Contemporary issues in thematic apperceptive methods* (pp. 3–40). Springfield, IL: Charles C. Thomas.
- Holt, R. R. (1978). *Methods in clinical psychology: Vol. 1. Projective assessment*. New York: Plenum.
- Holt, R., & Luborsky, L. (1958). *Personality patterns of psychiatrists: A study of methods for selecting residents*. New York: Basic Books.
- Holtzman, W. H. (1981). Holtzman Inkblot Technique. In A. I. Rabin (Ed.), *Assessment with projective techniques* (pp. 47–83). New York: Springer.
- Holtzman, W. H., Thorpe, J., Swartz, J., & Heron, E. (1961). *Inkblot perception and personality: Holtzman Inkblot Technique*. Austin, TX: University of Texas Press.
- Horowitz, M. J. (1991). *Person schema and maladaptive interpersonal patterns*. Chicago: University of Chicago Press.
- Ingram, R. E. (Ed.) (1986). *Information processing approaches to clinical psychology*. New York: Academic Press.
- Isen, A. M. (1984). Toward understanding the role of affect in cognition. In R. S. Wyer, Jr. & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 3, pp. 179–236). Hillsdale, NJ: Erlbaum.
- Isen, A. M. (1987). Positive affect, cognitive processes, and social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 203–253). New York: Academic Press.
- Johnston, M. H., & Holzman, P. S. (1979). *Assessing schizophrenic thinking*. San Francisco: Jossey-Bass.
- Kahana, B. (1978). The use of projective techniques in personality assessment of the aged. In M. Storandt, I. Siegler, & M. Elias (Eds.), *The clinical psychology of aging*. New York: Plenum.
- Kahill, S. (1984). Human figure drawings in adults: An update of the empirical evidence, 1962–1982. *Canadian Psychology*, 25, 269–292.
- Karon, B. P. (1981). The Thematic Apperception Test (TAT). In A. I. Rabin (Ed.), *Assessment with projective techniques: A concise introduction* (pp. 85–120). New York: Springer.
- Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivations* (Vol. 15). Lincoln, NE: University of Nebraska Press.
- Kelly, G. A. (1958). The theory and technique of assessment. In P. R. Farnsworth & Q. McNemar (Eds.), *Annual Review of Psychology*, 9, 323–352. Palo Alto: Annual Reviews.
- Kenny, D. T. (1964). Stimulus functions in projective techniques. In B. A. Maher (Ed.), *Progress in experimental personality research* (pp. 285–354). New York: Academic Press.
- Kenny, D. T., & Bijou, S. W. (1953). Ambiguity of pictures and extent of personality factors in fantasy responses. *Journal of Consulting Psychology*, 17, 283–288.
- Kihlstrom, J. F. (1984). Conscious, subconscious and preconscious: A cognitive perspective. In K. S. Bower & D. Michenbaum (Eds.), *The unconscious reconsidered*. New York: Wiley.
- Kihlstrom, J. F. (1987). The cognitive unconscious. *Science*, 237, 1445–1452.
- Kihlstrom, J. F. (1990). The psychological unconscious. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 445–464). New York: Guilford Press.
- Kihlstrom, J. F., & Cantor, N. (1984). Mental representations of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 17) New York: Academic Press.
- Klopfer, W. G. (1981). Integration of projective techniques in the clinical case study. In A. I. Rabin (Ed.), *Assessment with projective techniques* (pp. 233–264). New York: Springer.
- Kochanska, G. (1993). Toward a synthesis of parental socialization and child temperament in early development of conscience. *Child Development*, 64, 325–347.
- Koestner, R., Weinberger, J., & McClelland, D. C. (1991). Task-intrinsic and social-extrinsic sources of arousal for motives assessed in fantasy and self-report. *Journal of Personality*, 59, 57–82.
- Koppitz, E. (1968). *Psychological evaluation of children's human figure drawings*. New York: Grune & Stratton.
- Koppitz, E. M. (1984). *Psychological evaluation of human figure drawings by middle school pupils*. New York: Grune & Stratton.
- Kuhl, J., & Beckmann, J. (Eds.) (1994). *Volition and personality*. Seattle, WA: Hogrefe & Huber.
- Lerner, P. (1992). Toward an experiential psychoanalytic approach to the Rorschach. *Bulletin of the Meninger Clinic*, 56, 451–464.
- Leve, R. M. (1995). *Child and adolescent psychotherapy: Process and integration*. Boston: Allyn & Bacon.
- Lewin, K. (1935). *A dynamic theory of personality*. New York: McGraw-Hill.
- Locraft, C., & Teglassi, H. (1997). Teacher rated empathic behavior and children's T.A.T. stories. *Journal of School Psychology*, 35, 217–237.
- Luborsky, L., & Crits-Christoph, P. (1990). *Understanding transference: The CCRT method*. New York: Basic Books.
- Lundy, A. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment*, 49, 141–145.
- Machover, K. (1949). *Personality projection in the drawing of the human figure*. Springfield, IL: Charles C. Thomas.
- MacLeod, C., & Cohen, I. L. (1993). Anxiety and the interpretation of ambiguity: A text comprehension study. *Journal of Abnormal Psychology*, 102, 238–247.
- Maloney, M. P., & Ward, M. P. (1976). *Psychological assessment: A conceptual approach*. New York: Oxford University Press.
- McAdams, D. P. (1990). Unity and purpose in human lives: The emergence of identity as a life story. In A. I. Rabin, R. Zucker, R. Emmons, & S. Frank (Eds.), *Studying persons and lives* (pp. 148–200). New York: Springer.
- McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality*, 63, 365–396.
- McClelland, D. C., & Koestner, R. (1992). The achievement motive. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 143–152). New York: Cambridge University Press.
- McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review*, 96, 690–702.
- McGreevy, J. C. (1962). Interlevel disparity and predictive

- efficiency. *Journal of Projective Techniques*, 26, 80–87.
- Meissner, W. W. (1974). Differentiation and integration of learning and identification in the developmental process. *Annual of Psychoanalysis*, 2, 181–196.
- Meissner, W. W. (1981). Internalization in psychoanalysis. *Psychological Issues Monograph*, 50. New York: International Universities Press.
- Mendez, M. F., Ala, T., & Underwood, K. L. (1992). Development of scoring criteria for the clock drawing task in Alzheimer's disease. *Journal of the American Geriatric Society*, 40, 1095–1099.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Moffit, K. H., & Singer, J. A. (1994). Continuity in the life story: Self-defining memories, affect, and approach/avoidance personal strivings. *Journal of Personality*, 62, 21–43.
- Morgan, C. O., & Murray, H. A. (1935). A method for investigating fantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry*, 34, 289–306.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Murray, H. A. (1943). *Thematic Apperception Test manual*. Cambridge, MA: Harvard University Press.
- Murstein, B. I. (1965). The stimulus. In B. Murstein (Ed.), *Handbook of projective techniques*. New York: Basic Books.
- Murstein, B. I. (1968). Efforts of stimulus, background, personality, and scoring system on the manifestation of hostility on the TAT. *Journal of Consulting and Clinical Psychology*, 32, 355–365.
- Norcross, J. C. (Ed.) (1986). *Handbook of eclectic psychotherapy*. New York: Brunner/Mazel.
- Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.
- Piotrowski, C. (1984). The status of projective techniques: Or, "wishing won't make it go away." *Journal of Clinical Psychology*, 40(6), 1495–1502.
- Piotrowski, C., & Keller, J. W. (1989). Psychological testing in outpatient mental health facilities: A national study. *Professional Psychology: Research and Practice*, 20, 423–425.
- Piotrowski, C., Sherry, D., & Keller, J. W. (1985). Psychodiagnostic test usage: A survey of the Society for Personality Assessment. *Journal of Personality Assessment*, 49, 115–119.
- Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD clinical psychology programs. *Journal of Personality Assessment*, 61, 394–405.
- Plomin, R. (1986). *Development, genetics, and psychology*. New York: Erlbaum.
- Quinn, N., & Holland, D. (1987). *Cultural models in language and thought*. New York: Cambridge University Press.
- Rabin, A. I. (1981). Projective methods: A historical introduction. In A. I. Rabin (Ed.), *Assessment with projective techniques* (pp. 1–22). New York: Springer.
- Rappaport, D., Gill, M., & Schafer, R. (1968). *Diagnostic psychological testing* (Rev. ed.). New York: International University Press.
- Raynor, J. D., & McFarlin, D. B. (1986). Motivation and the self system. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (pp. 315–349). New York: Guilford Press.
- Ritzler, B. A., Sharkey, K. J., & Chudy, J. F. (1980). A comprehensive projective alternative to the TAT. *Journal of Personality Assessment*, 44, 358–362.
- Roback, H. B. (1968). Human figure drawings: Their utility in the clinical psychologist's armamentarium for personality assessment. *Psychological Bulletin*, 70, 1–19.
- Rogers, R. (Ed.) (1997). *Clinical assessment of malingering and deception*. New York: Guilford Press.
- Rubin, D. C. (Ed.) (1986). *Autobiographical memory*. New York: Cambridge University Press.
- Rummelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schematic and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rummelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.
- Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology*, 73, 482–486.
- Saltz, G., & Epstein, S. (1963). Thematic hostility and guilt responses as related to self-reported hostility, guilt, and conflict. *Journal of Abnormal and Social Psychology*, 67, 469–479.
- Sandler, J., & Rosenblatt, B. (1962). The concept of the representational world. *The Psychoanalytic Study of the Child*, 17, 128–145.
- Schank, R. C. (1990). *Tell me a story: A new look at real and artificial memory*. New York: Charles Scribner's Sons.
- Shapiro, D. (1965). *Neurotic styles*. New York: Basic Books.
- Shneidman, E. S. (1951). *Thematic test analysis*. New York: Grune and Stratton.
- Singer, J. L. (1981). Research applications of projective methods. In A. I. Rabin (Ed.), *Assessment with projective techniques* (pp. 297–331). New York: Springer.
- Singer, J. L., & Salovey, P. (1991). Organized knowledge structures and personality. In M. J. Horowitz (Ed.), *Personal schemas and maladaptive interpersonal patterns* (pp. 33–80). Chicago: University of Chicago Press.
- Singer, J. A., & Singer, J. L. (1994). Social-cognitive and narrative perspectives on transference. In J. M. Masling & R. F. Bornstein (Eds.), *Empirical perspectives on object relations theory*. Washington, DC: American Psychological Association.
- Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.
- Smith, B. L. (1994). Object relations theory and the integration of empirical and psychoanalytic approaches to Rorschach interpretation. In I. B. Weiner (Ed.), *Rorschachiana XIX: Yearbook of the International Rorschach Society* (pp. 61–77). Seattle, WA: Hogrefe & Huber.
- Smith, C. P. (Ed.) (1992a). *Motivation and personality: Handbook of thematic content analysis*. New York: Cambridge University Press.
- Smith, C. P. (1992b). Reliability issues. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 126–139). New York: Cambridge University Press.
- Snow, R. E. (1974). Representative and quasi-representative designs for research on teaching. *Review of Educational Research*, 44, 265–291.
- Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement. *Psychological Bulletin*, 112, 140–154.
- Stein, D. J., & Young, J. E. (Eds.) (1992). *Cognitive science and clinical disorders*. San Diego, CA: Academic Press.
- Stein, M. J. (1955). *The Thematic Apperception Test* (Rev. ed.). Cambridge, MA: Addison-Wesley.
- Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.
- Stinson, C. H., & Palmer, S. E. (1991). Parallel and distributed processing models of person schemas and psychopathologies. In M. J. Horowitz (Ed.), *Personal schemas and maladaptive interpersonal patterns* (pp. 339–377). Chicago: University of Chicago Press.
- Stricker, G., & Healey, B. J. (1990). Projective assessment

- of object relations: A review of the empirical literature. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 2, 219–230.
- Swensen, C. H. (1968). Empirical evaluations of human figure drawings: 1957–1966. *Psychological Bulletin*, 70, 20–44.
- Swindell, C. S., Holland, A. L., Fromm, D., & Greenhouse, J. B. (1988). Characteristics of recovery of drawing ability in left and right brain-damaged patients. *Brain and Cognition*, 7, 16–30.
- Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social Cognition. The Ontario symposium on personality and social psychology*. Hillsdale, NJ: Erlbaum.
- Teglasi, H. (1993). *Clinical use of story telling: Emphasizing the TAT with children and adolescents*. Needham Heights, MA: Allyn & Bacon.
- Thomas, A. D., & Dudek, S. Z. (1985). Interpersonal affect in Thematic Apperception Test responses: A scoring system. *Journal of Personality Assessment*, 49, 30–36.
- Tomkins, S. S. (1947). *Thematic Apperception Test*. New York: Grune & Stratton.
- Tomkins, S. S. (1979). Script theory: Differentiated magnification of affects. In H. E. Howe, Jr. & R. A. Dienstbier (Eds.), *Nebraska Symposium on Motivation* (Vol. 26). Lincoln, NE: University of Nebraska Press.
- Tomkins, S. S. (1987). Script theory. In J. Aronoff, A. J. Rabin, & R. A. Zucker (Eds.), *The emergence of personality* (pp. 147–216). New York: Springer.
- Tomkins, S. S. (1991). *Imagery, affect, consciousness* (Vol. 3). New York: Springer.
- Torgesen, J. K. (1977). The role of non-specific factors in task performance of learning disabled children: A theoretical assessment. *Journal of Learning Disabilities*, 10, 27–35.
- Tunis, S. L. (1991). Causal explanations in psychotherapy: Evidence for Target-and-Domain-Specific schematic patterns. In M. J. Horowitz (Ed.), *Personal schemas and maladaptive interpersonal patterns* (pp. 261–276). Chicago: University of Chicago Press.
- Uleman, J. S., & Bargh, J. A. (Eds.) (1989). *Unintended thought*. New York: Guilford Press.
- Urban, H. M. (1963). *The Draw-A-Person*. Los Angeles: Western Psychological Services.
- Vane, J. R., & Guarnaccia, V. J. (1989). Personality theory and personality assessment measures: How helpful to the clinician? *Journal of Clinical Psychology*, 45, 5–19.
- Veroff, J. (1992). In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 100–109). New York: Cambridge University Press.
- Watkins, C. E., Campbell, V. L., & McGregor, P. (1988). Counseling psychologists' uses of and opinions about psychological tests: A contemporary perspective. *The Counseling Psychologist*, 16, 476–486.
- Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology Research and Practice*, 26, 54–60.
- Weiner, I. B. (1977). Approaches for Rorschach validation. In M. Rickers-Ovsiankina (Ed.), *Rorschach psychology*. Huntington, NY: Krieger.
- Weiner, I. B. (Ed.) (1994). *Rorschachiana XIX: Yearbook of the International Rorschach Society*. Seattle, WA: Hografe & Huber.
- Wertheim, E. H., & Schwartz, J. C. (1983). Depression, guilt, and self-management of pleasant and unpleasant events. *Journal of Personality and Social Psychology*, 45, 884–889.
- Westen, D. (1991). The clinical assessment of object relations using the TAT. *Journal of Personality Assessment*, 56, 56–74.
- Westen, D. (1993). Social cognition and social affect in psychoanalysis and cognitive psychology: From regression analysis to analysis of regression. In J. W. Barron, M. N. Eagle, & D. L. Wolitzky (Eds.), *Interface of psychoanalysis and psychology* (pp. 375–388). Washington, DC: American Psychological Association.
- Willock, B. (1992). Projection, transitional phenomena, and the Rorschach. *Journal of Personality Assessment*, 59, 99–116.
- Wilson, A. (1988). Levels of depression and clinical assessment. In H. D. Lerner & P. M. Lerner (Eds.), *Primitive mental states and the Rorschach* (pp. 441–462). Madison, CT: International Universities Press.
- Wyatt, F. (1947). The scoring and analysis of the Thematic Apperception Test. *Journal of Psychology*, 24, 319–330.
- Wyer, R. S., Jr., & Srull, T. K. (Eds.) (1994). *Handbook of social cognition* (2nd ed., Vols. 1 & 2). Hillsdale, NJ: Erlbaum.

4.17

Computer Assisted Psychological Assessment

GALE H. ROID and W. BRAD JOHNSON
George Fox University, Newberg, OR, USA

4.17.1 INTRODUCTION	502
4.17.1.1 <i>Definitions and Distinctions</i>	502
4.17.1.2 <i>Brief History of Computer Assisted Psychological Assessment</i>	503
4.17.2 A TYPOLOGY OF COMPUTER ASSISTED PSYCHOLOGICAL ASSESSMENTS	504
4.17.2.1 <i>Test Administration</i>	505
4.17.2.2 <i>Computer Scoring</i>	506
4.17.2.3 <i>Descriptive Interpretation</i>	507
4.17.2.4 <i>Narrative Interpretation</i>	507
4.17.2.5 <i>Statistical–Actuarial Programs</i>	507
4.17.3 COMPUTER ASSISTED PSYCHOLOGICAL ASSESSMENT: ADVANTAGES	509
4.17.3.1 <i>Improved Administration and Scoring</i>	509
4.17.3.2 <i>Objectivity</i>	509
4.17.3.3 <i>Speed</i>	509
4.17.3.4 <i>Reliability</i>	509
4.17.3.5 <i>Cost Effectiveness</i>	509
4.17.3.6 <i>Expert Consultation</i>	510
4.17.3.7 <i>Flexibility</i>	510
4.17.4 COMPUTER ASSISTED PSYCHOLOGICAL ASSESSMENT: DISADVANTAGES	510
4.17.4.1 <i>Excessive Generality: The Barnum Effect</i>	510
4.17.4.2 <i>Lack of Validity</i>	511
4.17.4.3 <i>Depersonalizing the Assessment Process</i>	511
4.17.4.4 <i>Potential for Misuse and Client Harm</i>	511
4.17.4.5 <i>Computer as Clinician</i>	511
4.17.5 ETHICAL ISSUES	512
4.17.5.1 <i>Test Development</i>	512
4.17.5.2 <i>Basis for Scientific and Professional Judgments</i>	512
4.17.5.3 <i>Describing the Nature of Psychological Services</i>	513
4.17.5.4 <i>Competence</i>	513
4.17.5.5 <i>Professional Context</i>	514
4.17.5.6 <i>CAPA with Special Populations</i>	515
4.17.6 GUIDELINES FOR USERS OF COMPUTER-BASED TESTS AND INTERPRETATIONS	515
4.17.6.1 <i>Administration</i>	517
4.17.6.2 <i>Evaluation and Selection of CBTIs</i>	517
4.17.6.3 <i>Interpretation</i>	518
4.17.7 GUIDELINES FOR DEVELOPERS OF COMPUTER-BASED TEST SERVICES	519
4.17.7.1 <i>Human Factors</i>	519
4.17.7.2 <i>Psychometric Properties</i>	519
4.17.7.3 <i>Classification Strategy</i>	519
4.17.7.4 <i>Validity of Computer Interpretations</i>	519
4.17.7.5 <i>Facilitation of Review</i>	520

4.17.8 DIRECTIONS FOR THE FUTURE	520
4.17.9 CONCLUSIONS AND RECOMMENDATIONS	521
4.17.10 REFERENCES	522

4.17.1 INTRODUCTION

A common dilemma for clinical psychologists is the tension between the need for diagnosis and the complexities of the individual case. Although resources to help the client may be contingent on certain types of psychological disorders, concern for the client and the avoidance of stigmatizing labels clearly make assessment difficult when cases are unusual or complex. Also, time may be limited or resources of the client may be limited, making an extended assessment difficult to complete. In the midst of this dilemma, various methods of computer assisted psychological assessment (CAPA) have emerged and some have become prominent, especially with the advent of the personal computer in the 1980s, with both promise (e.g., Jackson, 1985) and potential problems (e.g., Matarazzo, 1986). This chapter surveys the definition, history, types of implementation, advantages and disadvantages, reliability and validity concerns, and issues of ethical and professional responsibility in the use of computers in assessment. No attempt will be made in this chapter to review or critique actual software or existing CAPA programs, except by example or reference, due to the rapidly changing nature of the technology and the continual updating of various programs. This chapter will, however, take a hard look at the proposed advantages of CAPA, and, instead of placing inordinate weight on the technical promise of computers, propose some firm limits on the use of CAPA. First, some definitions and important distinctions are made, followed by a brief history of the development of CAPA.

4.17.1.1 Definitions and Distinctions

CAPA is broadly defined as any application of computers to the development, administration, scoring or interpretation of tests, scales, inventories or questionnaires used in education or psychology.

Psychological texts such as Gregory (1996) have proposed similar definitions. Those appear consistent with early use of the term CAPA such as that employed by Fowler (1985). The present chapter will focus more narrowly on computer applications to cognitive or personality instruments in which administration, scoring, and interpretation are attempted. Methods of CAPA development and techniques of using

computers in test development are beyond the scope of the present chapter. For a review of some of the technical methods proposed for the development of computer-based psychological tests, see Green (1991a) on adaptive testing; Guastello and Rieke (1994) on the expert-system approach; Snyder, Widiger, and Hoover (1990) for a review of test-interpretive program development; and the review by Roid (1986).

By "assessment" we refer to the process of examining the entire range of information, including standardized tests, scales, inventories, questionnaires, protective tests, observations, interview data, histories, and other observations by experienced psychologists who study an individual for purposes of diagnosis, description, classification, treatment planning, pretherapy observation, or any other professional evaluation. Most psychometric textbooks make an important distinction between assessment and "testing" or "measurement." Testing usually involves the administration of standardized stimuli (e.g., test or inventory items) to clients whose responses are scored by objective or judgment-based scoring methods. Measurement is traditionally defined as the assignment of categories or scale values to such aspects of individuals as traits, attributes, attitudes, behavior, and preferences by a psychological testing instrument. The word "test" is typically reserved for tasks in which there are correct answers or problems to be solved, in contrast to "scale" or "inventory" or "questionnaire" which measure traits or preferences for which no single correct answer exists for all people.

Matarazzo (1986) made the most vivid distinction between testing and assessment by arguing that assessment should be reserved as a term to describe the process conducted by an experienced psychologist who gathers information for purposes of interpreting it and giving it meaning within the context of the examinee's total life history. Matarazzo's distinction would result in questions such as, "Can computers really provide assessments or only measurements which must be interpreted by professionals?" This penetrating question strikes at the heart of the ethical issues to be discussed in this chapter. Matarazzo was contrasting his definition of assessment with the technical, clerical, or computerized processing of test information. His position received immediate reactions such as that of Fowler and Butcher (1986) who argued for the view of the assessment process as

one that could include both clinical and computerized statistical assistance in the interpretive process.

The standards for psychological testing (American Psychological Association (APA), 1985) make a distinction between test administration and test interpretation, and warn that test developers should inform users of any special training or expertise needed for either administration or interpretation. Thus, CAPA may be used for test administration, for administration and scoring, or for a full range of administration and interpretation. Typically, a higher level of training is assumed for the interpretation of test results, such as graduate-level measurement or psychometrics courses, supervised assessment and reporting, and knowledge of concepts such as error of measurement. Also, a higher level of validation is required for any interpretations of tests that impact clients. The distinction between administration and interpretation of tests is critical to the evaluation of CAPA and the role of computers because it highlights the continuing responsibility of the psychologist to proactively interpret the results of the assessment process. Thus, we define another category of programs, computer-based test interpretive (CBTI) programs, to delineate those that emphasize interpretation, often with narrative descriptions of results.

A final distinction of great importance was first discussed in depth by Meehl (1954)—clinical versus statistical (actuarial) prediction. By clinical prediction, common usage would normally suggest the processing of information by the trained clinician who makes a prediction, diagnosis, classification, or evaluation of a client based on clinical experience, “intuition,” and professional judgment. By statistical prediction, common usage would suggest methods such as multiple regression or validated “cutting scores” being used to predict, diagnose, and classify an individual based on a relevant research database. Meehl’s (1954) lengthy discussion of the issues pointed out important distinctions such as the fact that both clinical judgment and statistical methods may be predicting an individual case from trends in previous group data—the group of all previous clients in the instance of clinical judgment. Later researchers such as Goldberg (1968) showed that clinical judgments are not necessarily more complex or “configural” as compared to statistical predictions, since simple linear equations effectively modeled the behavior of skilled judges who were assessing, for example, psychosis vs. neurosis from the Minnesota Multiphasic Personality Inventory (MMPI). It would be impossible to summarize

all of the careful distinctions among clinical vs. statistical methods, and the reader is referred to Meehl (1954), Goldberg (1968), or to the review by Garb (1994).

4.17.1.2 Brief History of Computer Assisted Psychological Assessment

Fowler (1985) and Moreland (1992) provide brief histories of CAPA. Highlights include the early attempts in the 1940s to computer score the Strong Vocational Interest Blank (SVIB), Meehl’s (1954) landmark book, and early 1960s versions of mainframe computer programs to score and interpret MMPI responses. To say that the history of CAPA has followed the development of computers and scanning equipment is obvious, but several landmarks are instructive. Following the refinement of mainframe computers in World War II, and until approximately 1970, most efforts to study or implement CAPA were conducted on mainframe computers with several critical attributes: (i) access to the computer was typically limited to professional computer operators, (ii) programs were developed and operated in “batch” mode where little interaction occurred between the user and the computer—interactions were specified in advance with various control commands, and (iii) the responses of examinees had to be scanned by a separate document scanner or key entered. Thus, “dynamic” entry of data was not possible at most installations. The advanced development of high-capacity scanning equipment is often attributed to Lindquist and his colleagues at the University of Iowa (e.g., Flanagan & Lindquist, 1951), for use on various educational achievement tests, although similar developments were employed for the SVIB and the MMPI (Moreland, 1992).

In the 1970s (and, perhaps, earlier in experimental laboratories), “time sharing” computer systems proliferated that connected the user via Teletype machine or early versions of computer terminals. Some CAPA applications were developed on systems initially designed for computer-assisted instruction. These “real time” systems allowed for a degree of interaction between the programmer and the system, between the user and the output, and, in some experimental applications, between the examinee and the time-share computer (e.g., Klingler, Miller, Johnson, & Williams, 1977). Klingler et al. developed an automated assessment system for psychiatric inpatients at a veterans administration hospital in Utah. The potential cost-benefit of such applications was immediately apparent, and this stimulated discussions at psychological conventions about

the ethical issues of on-line test administration and scoring.

Another important innovation in CAPA was the development of multistage “branching” or “adaptive” tests, which emerged from early psychometric studies in educational measurement (e.g., Angoff & Huddleston, 1958; Linn, Rock, & Cleary, 1969; Lord, 1968, 1971), from sequential methods in statistics (e.g., Cowden, 1946; Wald, 1947), and from the development of item-response theory (e.g., Birnbaum, 1968; Rasch, 1980). As early as the late 1960s, there was a considerable unpublished literature on adaptive testing methods and their application to psychological scales as well as educational tests (e.g., Bayroff & Sealy, 1967; Patterson, 1962; Roid, 1969). However, the first widely distributed applications of computerized adaptive testing emerged in the 1980s. Weiss (1983) was a key developer of methodology and applications in college and military settings. Operational programs on personal computers emerged in aptitude testing (e.g. McBride, 1988) and in conjunction with a large-scale project to computerize the Armed Services Vocational Aptitude Battery (ASVAB; Green, 1991a; Sands & Gade, 1983).

With the development of the first Apple computers and the release of the first IBM personal computer in approximately 1980, the widespread implementation and practicality of CAPA for the local psychologist was finally available. The early 1980s witnessed a flurry of rapid development and distribution of test scoring and interpretive programs for microcomputers (e.g., Roid & Gorsuch, 1984). In reaction to the swift proliferation of CAPA programs, Matarazzo (1983) published a stern warning about the lack of validation of narrative interpretations and the potential dangers of distribution to inexperienced users. Special series of articles appeared in the *Journal of Consulting and Clinical Psychology* (Butcher, 1985), and *Computers in Human Behavior* (Mitchell & Kramer, 1985.) These early articles were a mixture of praise about the potential of the methodology and warnings about the limitations and ethical consequences of irresponsible usage. In response to this outpouring of attention to CAPA, the APA (1986) published a booklet of guidelines for development and usage of “computer-based tests and interpretations” (see Table 2). Several resource books, cataloging various programs and user options, were published (e.g., Butcher, 1987; Krug, 1984).

Because of the rapid innovation in computer technology, and, in the mid-1990s, the development of the world-wide web, Internet, video conferencing, and multimedia desk-top compu-

ters, it is difficult to anticipate the developments in CAPA. Certainly, the addition of multimedia to assessment instruments will be more easily achieved and affordable by more test developers. Dynamic video segments within standardized tests could make real-life situational assessment more feasible. In any case, the potential and the new ethical concerns stimulated by the prospect of test processing by Internet or e-mail services looms large on the horizon.

In the next section, we review a typology of CAPA programs. Included is a review of some of the key literature that documents the current status of CAPA and further definitions and distinctions.

4.17.2 A TYPOLOGY OF COMPUTER ASSISTED PSYCHOLOGICAL ASSESSMENTS

Not all applications of CAPA have the same level of complexity or developmental sophistication to support them. Thus, it is important to have a typology of various programs so that the proper role of CAPA can be evaluated. It will be the overall recommendation of this chapter that the psychologist carefully evaluate the level and type of CAPA product considered for clinical use, and perhaps, “draw a line” of restriction in terms of the type, technical quality, and validity of CAPA results actually employed in the evaluation of clients. For these reasons, a typology of CAPA programs, derived from the literature and the previous work of the senior author (Roid & Gorsuch, 1984; Roid, 1986), is presented below. First, a few key distinctions and definitions are discussed.

As stated earlier in this chapter, a wide range of computer “assistance” is used in the development of psychological tests and assessments, and we have chosen to delete these programs from our typology in favor of emphasis on CAPA products that may be directly used in client assessments. Also, there are a potentially large number of CAPA products available in the “employment testing” industry that are not considered due to the clinical focus of the current discussion. Finally, computer-based products that are designed for on-line delivery of psychotherapeutic exercises or cognitive training are excluded, even though some of them include assessment components.

Significant differences exist between CAPA programs for cognitive, neuropsychological, and personality assessments. Some of the same statistical techniques have been applied to the profile scores of both ability and personality batteries, but, for the most part, each has a particular type and style of presentation. For

example, it is now common in the software to score intelligence scales to include computations, frequencies, and probabilities of differences between two or more subtests or indexes (e.g., verbal IQ vs. performance IQ). Statistical contrasts of differences are more rare in personality assessment, where, instead, visual inspection of the profile of scores is more common. In computerized interpretive programs in the personality realm (e.g., Lachar, 1984), it is common to include “critical items” and correlates between clinician ratings and “profile elevations” (ranges of scores, such as 70–79T on the MMPI). However, in terms of the typology that follows, the collective statistical techniques can be classified together as “profile and statistical analysis.”

4.17.2.1 Test Administration

Table 1 presents the typology. Some software is designed to administer tests and to record the results, which may be printed or briefly summarized on the computer screen. Administration can include highly sophisticated, multimedia programs with attached database capability, such as the MicroCog by Powell et al. (1993), an on-line testing system to detect cognitive impairment in older adults. Programs can vary from basic “page turning” software that simulates paper-and-pencil tests, to those such as MicroCog that include graphic displays, reaction time, optional sound, and delayed memory trials also timed. Of particular importance in this category of the typology, are programs such as the Conners’ (1995) Continuous Performance Test (CPT), used to assess attentiveness, particularly for children and adults referred for attention-deficit evaluation. The CPT assesses visual vigilance in scanning and concentrating on a stimulus array—a type of performance that must have accurate, split-second timing of both stimuli and responses—a perfect application of computers in test administration.

A separate category of test administration is computer-adaptive testing (CAT). CAT requires specialized software to implement the sophisticated (usually item-response theory or Bayesian) statistical model on which it is based. Several prominent examples exist, including the newer versions of the computerized Graduate Record Exams (Educational Testing Service, 1995), the ASVAB (Sands & Gade, 1983), and, one of the first widely published commercially published tests, the Differential Aptitude Tests (DAT Adaptive) Computerized Adaptive Edition (Psychological Corporation, 1987). The essential attribute of CAT administration is that the difficulty of the test items is tailored

Table 1 A typology of computer assisted assessment.

<i>Test administration</i>
Conventional test administration
Multimedia or specialized test administration
Computerized adaptive testing (CAT)
<i>Computer scoring</i>
Conventional scoring
Statistical scoring and profile analysis
Graphic displays
Database analysis
<i>Descriptive interpretation</i>
<i>Narrative interpretation (CBTI programs)</i>
Informal narrative
Clinician-modeled or expert system
<i>Statistical-Actuarial interpretation</i>

dynamically (Lord, 1968) to the ability level of the examinee during the testing session. Research (e.g., Green, 1991a) has shown that modern implementations of CAT can reduce testing time by half while maintaining an acceptable level of measurement error. To implement such a system requires complex psychometric development in which each item in the pool of available items is given field trials and calibrated statistically. Programs such as MicroCAT (Assessment Systems, 1990) or other specialized computer programs are then used to implement the presentation of items. A broad range of “branching rules” and scoring methods are now available for such tests, depending on the type of item-response theory model used in the calibration of the items (e.g., Wainer, 1990).

While the statistical complexity of CAT applications may be beyond the level of psychometric training for many clinical psychologists, one can consult the professional reviews of such tests in the *Buros Mental Measurements Yearbooks*, or the current on-line electronic versions of these reviews. Alternatively, legitimate CAT programs should have an accompanying manual which can be evaluated for conventional reliability and validity data, just as any published test is evaluated. Psychological conventions and meetings often include exhibits or demonstrations of computer software, which can be a more cost-effective way to review these products, since publishers are often reluctant to distribute expensive software for review. As a final recourse, one should consult a psychometric psychologist at a local university who can help to evaluate the technical qualities of programs considered for purchase.

4.17.2.2 Computer Scoring

Conventional scoring software has typically processed answer sheets or key-entered item responses and printed basic raw scores and various derived scores such as percentiles or standard scores. For many years, such programs had a minimum of graphic display, due to the restricted nature of graphics possible until the advent of the high-speed laser printer. As mentioned in the brief history of CAPA, the earliest versions of conventional scoring were on mainframe computers for the SVIB or the MMPI. Many of these programs were quite creative and complex, relying on "typeface" characters to plot profiles or display histogram-type plots of scores.

At the next level in the typology are scoring programs that include sophisticated statistical computations and profile analysis. A proliferation of "scoring assistant" programs has emerged, such as the "compuscore" series for the Stanford-Binet and Woodcock-Johnson at Riverside Publishers, the PsyTest and Wechsler Intelligence Scale for Children-Third Edition (WISC-III) and Wechsler Individual Achievement Test (WIAT) programs of The Psychological Corporation, the Western Psychological Services Test Report series, the software systems developed by Psychological Assessment Resources, and the "Assist" software for the tests published by American Guidance Service, to name a few of the larger groupings of scoring programs. Although many of these programs also include interpretive functions, to be discussed in the next section, the scoring sections of these programs tend to be more graphic and more sophisticated than earlier, conventional scoring programs (Roid, 1985). For example, a much wider array of profile indexes, score-difference analyses, critical item comparisons, and profile-matching routines are included than previously available programs (with exceptions being some of the complex, early MMPI interpretive programs). Two examples, one in personality and one in cognitive assessment, may illustrate the characteristics of such programs. The Tennessee Self Concept Scale (Roid & Fitts, 1988) computerized scoring program includes a wide array of research-based profile indexes, checks on the validity of response patterns, faking-good scales, critical-item lists, and a multivariate profile matching method that is implemented on a complex, color-printed display. The profile-matching method, interval-banded profile analysis, was initially developed by Huba (1986) to provide a statistical test of the degree of match between prototypical profiles (stored within the computer program)

and those of new profiles being scored. The essential part of the matching routine is a multivariate chi-square test of "fit" between the profile pattern (which is allowed to vary within "bands" or ranges of scores) and "target" prototypical patterns that have been established through statistical studies of clinical and normative cases. A second example of sophisticated profile analysis is in the Wechsler Scoring Assistant (Psychological Corporation, 1992), where profile-score differences among all the subtests of the WISC-III are analyzed statistically, and the regression-prediction method of calculating ability vs. achievement discrepancies (WISC-III vs. Wechsler Individual Achievement Test scores) are included for screening of learning disabilities.

As more and more technical advances are made in computer printers, compact-disk and computerized video displays, the more sophisticated the graphic display of test profiles and score patterns will become. In the final level of the typology are programs that allow for the archiving of multiple cases, set up and usage of extensive databases of test results, and statistical manipulation of these data. Examples of such programs are found among the offerings of all the publishers of the large achievement-test batteries, who typically offer software for analyzing trends in achievement data across an entire school district. The most sophisticated versions of these programs allow for the assessment of growth or change at the individual, classroom, and school levels of analysis.

Based on the experience of the senior author of this chapter, who participated in the development of several computer-scoring systems (e.g., Roid, 1985; Roid & Fitts, 1988; and the Wechsler Scoring Assistant, Psychological Corporation, 1992), these profile-analysis programs are typically based on extensive data analysis, not informal or subjective processes that often occur in the "informal narrative" category of the typology. Most of the profile analyses are based on actual data from the standardization and validity studies of the major tests, and extensive staff, resources and time are invested in the development, and (elaborate) cross-checking of program accuracies. Because it is well known that clerical errors in scoring standardized tests are all too common, these scoring programs provide a valuable service to clinical psychology by delivering accurate scores. Further, the complexity of profile analysis, difference-score computation and analysis, and profile-pattern matching would not be feasible with hand-scoring methods. Except for a few unusual cases, the statistical scoring programs tend to be well documented in the technical manuals published with them.

Thus, in summary, the sophisticated “scoring assistant” model of computer scoring software has distinct advantages for the clinical psychologist. Time saved by the clinician in using scoring software could be reinvested in more time with the client or additional personal interview or case follow-up.

4.17.2.3 Descriptive Interpretation

A descriptive type of program would generate sentences of explanation, such as “The client has a significantly higher score on subtest three as compared to the other subtests in the profile,” along with the printed scores and profile. The distinction between this level of description and the narrative interpretations described in the next section of the typology is that description remains tied to the facts and does not indicate cause-effect relationships or connections with research or clinical findings. The descriptions would be analogous to the phrases used in the results section of a research article—reporting the findings before they are discussed or interpreted. As noted in the example above, the descriptions can be rooted in sophisticated statistical comparisons between scores (e.g., Silverstein, 1981).

Some of the first computer interpretive programs widely used in the 1980s had redundant, printed phrases that described score elevations for multiple scores, using exactly the same wording. The best examples of descriptive interpretation include sophisticated “sentence generators” that compose explanations using a variety of modifiers, sentence construction and style, similar to the variety present in good report writing. One example is that of the Barclay Classroom Assessment System (Barclay, 1983) which varied pronouns such as “he” and “she,” and used research on the scaling of verbal phrases (e.g., Lichtenstein & Newman, 1967; Pohl, 1981) to compose explanations of score elevations. For example, scale values for descriptors such as “fairly often” vs. “very infrequently” can be used to anchor score-level descriptions, in a way more precise than an informal or subjective use of such language. Use of more precise methods of description, could increase the potential of such programs to maintain a more objective level of description, that is, sticking to the facts.

4.17.2.4 Narrative Interpretation

The worst examples of informal, unvalidated, narrative interpretations of psychological tests were the target of Matarazzo (1983, 1986) in his critiques of CAPA. By informal narrative we

mean the wordy descriptions printed on computer reports of assessment results that attach clinical significance, often based on clinical lore or theoretical explanations, to the patterns or levels of scores. Such narratives have been faulted for lack of validation of specific sentences or phrases and for lack of attention to individual client variations. Informal narratives also have the impact of vague generalities that result in a “Barnum effect” in which enough truth lies within a complex of statements if one emphasizes the accurate parts and minimizes the inaccurate parts. In research on the accuracy of computerized reports, Adams and Shore (1976) found an inverse relationship between length of reports and their accuracy rated by clinicians. For these reasons, clinicians should be very cautious about the validity of informal narrative programs.

Basically, the key feature of such programs, that clinicians should be able to discern from the manual and advertising material published with the computer program, is whether or not the narrative sentences have been accumulated from empirical research or whether they were written by the author(s) of the program without validation studies. Another key is whether or not the statements were validated by empirical linking of clinician ratings and scores, where both were collected in the same research studies. Positive examples of proper validation of narrative programs are reviewed by Snyder et al. (1990). Published programs that include extensive empirical validation of narrative reports include Lachar (1984) and Snyder (1981), to name only two prominent examples.

By clinician-modeled programs we mean those that either (i) employ the process used by a renowned clinician within the logic of the computer program, or (ii) employ statistical models of the process used by expert clinicians, determined through research on clinical judgment (e.g., the methods described by Goldberg, 1968). Examples of the former include the WISC-III program of Kaufman (1996) that implements his documented approach to interpretation of the Weschler scales (Kaufman, 1994). Examples of the latter have never been implemented, as far as we know, although the methodology is particularly promising. For a review of expert systems and their application to CAPA, see Guastello and Rieke (1994).

4.17.2.5 Statistical-Actuarial Programs

The term “actuarial” as applied to psychological assessment was coined by Meehl (1954) in analogy to the actuary process in insurance-

risk determination. Gregory (1996) presented a classic definition of an actuarial interpretation, attributed to J. O. Sines, as founded on "the empirical determination of the regularities that may exist between specified psychological test data and equally clearly specified socially, clinically, or theoretically significant nontest characteristics of the person tested" (p. 579). Thus, an actuarial approach is databased and must show a statistical link between information collected outside the test (e.g., clinician's observations) and the test scores or patterns of test results. At their best, narrative statements appearing in a truly actuarial program would not be based on clinical opinion, but rather on rigorous research linking test and nontest information.

A classic example of an actuarial approach to the MMPI is provided by Marks and Seeman (1963) who defined a 4-8-2 profile "code type" as follows:

- (i) Scales 4, 8, and 2 over 70T
- (ii) Scale 4 minus 2 less than 15T
- (iii) Scale 7 not to exceed 4 by more than 4T
- (iv) Scale 8 minus 2 less than 15T
- (v) Scale 8 minus 7 more than 5T
- (vi) Scale 8 minus 9 more than 10T
- (vii) Scale 9 less than 70T
- (viii) Scales L and K less than F, F less than 80T

Thus, detailed specifications are given for the entire profile pattern, not just the high scores on 4 (psychopathic deviate), 8 (schizophrenia), and 2 (depression). Even the validity scales, L (lie scale), F (frequency), and the suppressor scale K, are used to verify the accuracy of the profile. Marks and Seeman (1963) reported that patients obtaining this profile were mainly diagnosed psychotic (71% schizophrenic, paranoid type) though some were seen as personality disorders (e.g., 21% sociopathic). Note that these are percentages for a given sample of clinical patients, and that they do not predict with 100% accuracy, as with all probability relationships. Therefore, the most accurate statement that should appear in narrative form would be something like, "some research studies of clinical patients have shown a frequency of about 70% schizophrenic-paranoid for this pattern of scores," and the clinician should be careful to screen all such comments to be sure they apply to the current case.

Another type of actuarial approach called "empirical correlates" was developed by Lachar (1984) for the Personality Inventory for Children (PIC) computer interpretations. Lachar administered clinical checklists to psychologists who had interviewed children who were subsequently tested with the PIC and, previously, known to have certain diagnoses. In addition, a

large sample of children in a normative sample were given the PIC and the checklist was completed by examiners for these children also. For certain scales (and certain scale elevations, e.g., scores greater than 79T), empirical correlates were those descriptive statements statistically associated with the "elevation" (score in the clinical range). For example, for the achievement scale, a percentage (e.g., 73%) of clinical cases would show symptoms such as behavioral adjustment difficulties or negative self concept (as indicated on the clinical checklist), when the achievement scale score exceeded 79T. To confirm such findings, the same correlational study would be repeated as a cross-validation. Since empirical findings always "shrink" upon cross validation, the percentage may reduce to 70% of cases in the above example.

The best statistical-actuarial interpretation programs are those that meet the standards suggested by Snyder et al. (1990), where cut-score rules, program logic, and narrative sentences have been subjected to empirical research and are well documented in a technical manual. Also, the best of such programs have been subjected to validation research that surveys clinical users of the reports, collects accuracy ratings, and assesses the impact on clinical decisions of various client reports in rigorous follow-up studies. The number of programs with such rigorous development are, unfortunately, few. Some systems that come close are the larger, well-researched MMPI programs, the PIC (Lachar, 1984), the Marital Satisfaction Inventory (Snyder, 1981), and the Wechsler programs developed by the Psychological Corporation (1994), to name only four examples. Even these programs can be faulted in that every phrase or descriptor may not have been subjected to empirical trials based on moderator variables, and all may not have been examined for report accuracy. Such research, as with all good construct validation, takes decades of accumulated research. One trend and positive aspect of some of the newer programs (e.g., WISC-III Writer, Psychological Corporation, 1994), is the provision for placing the clinician in control of the final collection of narrative statements that appear in the report. As will be discussed in later sections of this chapter, the individual clinician must maintain control and oversight over the selection and accuracy of all statements generated by computer interpretations that are employed in case reports.

With the typology of CAPA programs in mind, the following section reviews some of the advantages and disadvantages of CAPA. Key literature is cited in the next section.

4.17.3 COMPUTER ASSISTED PSYCHOLOGICAL ASSESSMENT: ADVANTAGES

Computerized approaches to psychological assessment have rapidly become standard practice in most mental health treatment settings. Not only have computer administration and scoring of various tests become commonplace, but CBTIs have become a booming industry in their own right. In 1989, a survey of 413 mental health facilities in the USA revealed that 53% of these major facilities employed some form of CBTI (Piotrowski & Keller, 1989). CAPA offers numerous advantages to users, organizations, and clients alike. The following are some of the more salient advantages.

4.17.3.1 Improved Administration and Scoring

Computer administration of psychological tests serves to enhance standardization and clinician control over the testing process. In addition to reduced time for administration and rapid availability of feedback, computer administration allows presentation of even complex testing stimuli (Krug, 1987) and early identification of examinees who misunderstand directions or test stimuli (Wise & Plake, 1990). Though some have expressed concerns about the equivalence of computer administered test scores, reviews suggest that when instructions are similar, computer-administered tests generally yield scores which are equivalent to paper and pencil versions (Finn & Butcher, 1991). Additionally, preliminary research indicates that computer-administered tests are not only acceptable to examinees but often preferred over conventional testing procedures (Burke & Normand, 1987; Finn & Butcher, 1991). Specifically, examinees report greater interest, less anxiety, and greater comfort in responding to computer generated test stimuli. Computer-administered tests are also quite useful for more disturbed clinical populations who, by virtue of their level of disorganization or inattention, may respond more effectively to a computer than another person (Bloom, 1992).

4.17.3.2 Objectivity

In his early call for a good actuarial “cook-book” for use in psychological assessment, Meehl (1956) noted the potential benefit of decreased distortion and bias, in the recording, storage, and retrieval of test data and interpretive material. Most appear to agree that CAPA can substantially reduce errors associated with ignorance, bias, or stereotyping on

the part of clinicians (Butcher, 1987; Dahlstrom, 1993). Once interpretive rules are developed and programmed, they are automatically applied to protocols regardless of extraneous circumstances.

4.17.3.3 Speed

Perhaps the most pragmatic and obvious advantages of computer assisted psychological assessment is the potential for marked reduction in time required for administration, scoring and interpretation of psychological instruments (Butcher, 1987; Kleinmuntz, 1969, 1975; Wise & Plake, 1990). Administration time for both achievement and personality tests is significantly reduced by computer administration (Wise & Plake, 1990). Computerized scoring and interpretive report generation serve to enhance timely processing of test data by clinicians and subsequent delivery of feedback to consumers. In addition, these substantial reductions in time are associated with improvements both in consistency and accuracy.

4.17.3.4 Reliability

Computer assisted scoring and interpretation of psychological tests serves to radically reduce error variance. Butcher (1987) noted “The computer seldom has an off day as human test interpreters do” (p. 5). As a result, well developed test interpretation programs offer nearly perfect reliability for the functions of scoring and compiling preprogrammed interpretive statements (Burke & Normand, 1987; Graham, 1993; Krug, 1987). Once scale correlates have been identified and replicated, they can easily be stored and reliably recalled when particular scale elevations or score configurations appear in a protocol.

4.17.3.5 Cost Effectiveness

Computer scoring and interpretation of psychological tests generally result in marked reduction in clinician time and, therefore, expense to the consumer (Burke & Normand, 1987; Butcher, 1987; Graham, 1993). Though computerized systems will certainly require an initial outlay for equipment, software, and training, long-term savings to both users and consumers may be expected. With the advent of managed behavioral healthcare and increasing attention to issues such as cost-effectiveness, efficiency, and treatment utility, CAPA offers several ways for clinicians to reduce the time and expense required for a thorough assessment.

4.17.3.6 Expert Consultation

Many modern day CBTI programs appear to have effectively realized Meehl's (1956) hope for systems which employ large databases and widely representative samples on which to base predictions and diagnoses. Many well developed CBTIs systematically organize and access massive normative databases and extensive bodies of empirical research findings to bolster and undergird test interpretations (Krug, 1987). In spite of some methodological weaknesses, research on the equivalency of computer generated vs. clinician generated test interpretation suggests many CBTIs perform at least as well as clinicians (Bloom, 1992; Burke & Normand, 1987; Finn & Butcher, 1991). For example, Kleinmuntz (1969), in a large scale multi-site study, found that computer generated MMPI interpretations performed as well as expert MMPI interpreters and surpassed the performance of average clinicians in predicting a client's primary clinical problem.

Computer generated interpretations may ideally offer the skilled clinician a source of expert consultation. Due to the voluminous information available via the computer system, interpretations generated by a CBTI may markedly enhance both the accuracy and comprehensiveness of clinical decision making, treatment planning, and feedback to the client. Garb (1994) pointed out that even when a CBTI report appears to conflict with interview, historical, and observational data, the report can still be quite valuable in leading the clinician to consider alternative hypotheses and collect additional data. CBTIs may be most helpful when a case is ambiguous and the diagnosis unclear. Also, as a result of the "expert" and "objective" look and sound of computerized psychological reports, CBTIs may be particularly advantageous in forensic settings (Butcher, 1987) where such reports may be viewed as a form of outside and corroborating opinion. Finally, computer generated reports are particularly advantageous as a source of objective and expert opinion in the psychotherapy enterprise (Finn & Butcher, 1991). Here the test interpretive report is presented as a consultation by an outside expert and introduced for the purpose of clarification and discussion.

4.17.3.7 Flexibility

A final advantage of computer assisted psychological assessment is the flexibility possible in such systems (Graham, 1993; Kleinmuntz, 1975). Perhaps the most tangible example of this flexibility is the development

and ongoing refinement of adaptive or tailored testing programs. Here the system is programmed to adjust the difficulty level or volume of material in discrete areas to an individual examinee. Examples include research with the ASVAB and the MMPI. Sympson, Weiss, and Ree (1982) found that an adaptive version of the ASVAB produced validity coefficients which were equivalent to conventional administration of the test in spite of the fact that the adaptive versions were typically one-half the length of their conventional counterparts. Similarly, Roper, Ben-Porath, and Butcher (1991) found that among 155 college age subjects, an adaptive version of the MMPI-2 (averaging 28% shorter in length) produced profiles which were equivalent to conventional MMPI-2 administration.

4.17.4 COMPUTER ASSISTED PSYCHOLOGICAL ASSESSMENT: DISADVANTAGES

In addition to the many advantages inherent in CAPA, there are also concerns and unresolved dilemmas in the development, implementation, and utilization of computerized systems. Below, are highlighted several of these potential disadvantages. These include the problem of excessive generality, scanty validity, potential for depersonalizing the assessment process, potential for misuse or client harm, and the danger inherent in viewing the computer as a competent clinician.

4.17.4.1 Excessive Generality: The Barnum Effect

Computer generated testing interpretive reports have been criticized soundly for their frequently broad and highly generalized narrative statements (Butcher, 1987; Groth-Marnat & Schumaker, 1989; Matarazzo, 1986). Butcher described this as the problem of excessive generality, otherwise known as the Barnum effect or the Aunt Fanny report. CBTIs are notorious for offering patient descriptions based on insufficient empirical research. These narrative statements then apply to most human beings and are merely modal statements vs. person-specific descriptors.

This problem appears to have consistent empirical demonstration. O'Dell (1972) found that subjects perceived Barnum reports to be more accurate descriptions of their own personality than actual computer generated interpretations of their MMPI profiles. O'Dell concluded that statements with very high base rates tend to be believed or concurred with. Guastello, Guastello, and Craft (1989) found

that 58% of respondents rated Barnum reports for the Personality Profile Compatibility Questionnaire as quite accurate descriptions. Nonetheless, actual CBTI reports for this measure were rated as significantly more accurate than the Barnum reports. Less encouraging was a similar study employing a CBTI for the Exner Rorschach system. Prince and Guastello (1990) reported that the Exner report offered only 5% discriminating power for any one patient when compared with bogus Exner interpretive reports. Perhaps more concerning was the finding that approximately 60% of the CBTI statements contained statements which merely described characteristics of the outpatient population in general. Related to this concern regarding Barnum statements and excessive generality is the frequently expressed concern regarding an "aura" of credibility or accuracy potentially attributed to CBTI reports by consumers. The scanty research in this area suggests that this concern may be largely unfounded and consumers rate the quality, credibility, and accuracy of CBTI or clinician generated reports as essentially similar (Andrews & Gutkin, 1991).

4.17.4.2 Lack of Validity

One of the most glaring problems inherent in the development and use of CBTI report systems is the pervasive dearth of empirical evidence of their validity. Very few of the existing CBTI systems have been validated in even a rudimentary manner (Butcher, 1987; Finn & Butcher, 1991; Matarazzo, 1986). Rather than actuarial programs based exclusively on empirically derived base rate data, most CBTI reports are based on clinical lore or the conclusions and hypotheses of expert clinicians (Gregory, 1996; Groth-Marnat & Schumaker, 1989). Concerns regarding validity are often expressed as part of a larger concern about the wide range in product quality within the CBTI market. Butcher (1987) wondered at the "mind-boggling" (p. 6) array of computers and software packages from which clinicians might choose. While some systems are based on reasonably rigorous development procedures, most are not. As a result, many psychologists remain quite skeptical of CBTI programs (Burke & Normand, 1987) and avoid using them.

4.17.4.3 Depersonalizing the Assessment Process

Critics of computerized assessment systems often express concern that introduction of

computers in the assessment process will serve to heighten the distance between the clinician and client and lead to a more sterile and perhaps dehumanizing assessment process (Burke & Normand, 1987; Krug, 1987). In fact, research does not substantiate this concern and instead, suggests quite the opposite. Most people appear to prefer computer administered tests and some are more truthful in response to computer administered questions (Fowler, 1985). Computer administered assessments appear particularly useful for more disturbed or anxious clients.

4.17.4.4 Potential for Misuse and Client Harm

Later in this chapter we will address a range of potential ethical problems inherent in the use of computer assisted assessment. The potential for misuse of CBTIs and resulting harm to clients is substantial and results from difficulties with their current use (Butcher, 1987; Groth-Marnat & Schumaker, 1989). First, because CBTI programs and services are widely available and sold to a wide range of professionals, it is likely that professionals without adequate awareness of the limitations of CBTIs will apply them to clients regardless of context or important mitigating factors. A related concern involves the potential for factors indigenous to computerized assessment, but irrelevant to the purposes of the test, to significantly alter test performance (Moreland, 1985). Second there may be a tendency for consumers to uncritically accept statements generated by a computer as more factual than those generated by a clinician. Finally, because CBTI reports are rarely signed, there are serious concerns regarding professional responsibility and legal culpability (Gregory, 1996). Without a qualified psychologist assuming responsibility for the service offered, potential for misuse is enhanced.

4.17.4.5 Computer as Clinician

The final potential disadvantage in employment of computers in the assessment process has to do with the danger that well trained psychologists might lose control (Krug, 1987) of the assessment process. Specifically, this would be a loss to mechanization, technology, and large marketing interests. Though clinicians clearly stand to benefit from the speed and reliability of computer driven assessments, there is concern that the expert clinician of the past will become the testing technician of the future (Butcher, 1987; Butcher, Keller, & Bacon, 1985; Groth-Marnat & Schumaker, 1989; Matarazzo, 1990). As a result of the highly professional

appearance of many CBTI reports, both users and recipients of computerized narratives may confuse them with comprehensive assessments. To the extent that the skills of the human clinician are relegated to a position of diminished importance in generating assessment outcomes, and to the extent that clinicians perceive themselves as less responsible for these outcomes, there certainly exists greater risk to the profession and the consumer.

4.17.5 ETHICAL ISSUES

In light of the rapid proliferation of CAPA techniques, and CBTIs in particular, it is not surprising that this burgeoning area of research and practice has become the focus of a wide range of ethical concerns. The APA's Ethical Principles and Code of Conduct (APA, 1992) has as its primary goal "the welfare and protection of the individuals and groups with whom psychologists work" (p. 1599). Principle C from the APA code is perhaps most relevant to the ethical issues inherent in computerized assessment, "Psychologists uphold professional standards of conduct, clarify their professional roles and obligations, accept appropriate responsibility for their behavior and adapt their methods to the needs of different populations" (p. 1599). With these goals in view, we will consider several of the most salient ethical obligations for psychologists involved in CAPA. If neglected, each could serve as a source of potential harm to consumers of computerized services.

4.17.5.1 Test Development

The Ethics Code (APA, 1992) states clearly that psychologists involved in the development and provision of computerized assessment services accurately describe the purpose, development procedures, norms, validity, reliability, and applications of the service as well as particular qualifications or skills required for their use. Psychologists participating in such product development should attempt to clearly link interpretive statements to specific client scores or profiles, qualify narrative statements to minimize the potential for misinterpretation and perhaps provide some form of warning statement to alert users to the potential for misinterpretation (Hofer & Green, 1985). At the very least, the developer might note that the clinical interpretations offered in narrative printouts are not to serve as the sole basis on which important clinical decisions are made (Matarazzo, 1986). Adherence to the highest standard of the profession would also

require developers to provide rather detailed information regarding the system's development and structure in a separate manual. Because individual users are responsible for determining the validity of any CBTI for individual test-takers, availability of such system information is critical. Bersoff and Hofer (1991) noted that in spite of the apparent conflict between the developer's proprietary interest in the product and the clinician's need to responsibly evaluate the service, open and critical review of tests and CBTIs is critical for ensuring the quality of such materials and upholding the profession's ethical code.

4.17.5.2 Basis for Scientific and Professional Judgments

Psychologists rely on scientifically and professionally derived knowledge when making both scientific and professional judgments. The Ethics Code (APA, 1992) also states explicitly that "Psychologists select scoring and interpretation services (including automated services) on the basis of evidence of the validity of the program and procedures as well as on other appropriate considerations" (p. 1604). While psychologists are clearly compelled to justify their professional statements and behavior with empirically derived evidence, the current state of CBTI development makes this a difficult task indeed. The overwhelming majority of automated interpretive programs lack even preliminary forms of established validity (Lanyon, 1984). With this concern in mind, Matarazzo (1986) insisted that until research establishes (even the most primitive) validity for CBTIs, "It is essential that they be used only as tools by the clinician trained in their use and not as equivalents of, and thus substitutes for, professional education and training" (p.14).

Empirical validation of CBTI systems is exceptionally difficult given the exhaustive range of potential test scores and profiles. As a result, there are no purely actuarial interpretive programs in existence (Butcher et al, 1985; Fowler, 1985). Instead, most CBTIs offer a form of automated clinical prediction (Graham, 1993) in which published research, clinical hypotheses, and clinical experience on the part of an "expert" clinician are integrated into interpretive narrative statements. Nonetheless, the validity of such statements cannot be assumed by users of such reports and must be demonstrated every bit as much as the validity of the test on which it is based. Psychologists who employ CBTI reports must clearly understand the basis for the statements offered by such services, their validity or lack thereof, and

take reasonable steps to ensure that those with whom they work are not harmed by the irresponsible or uncritical use of such reports.

4.17.5.3 Describing the Nature of Psychological Services

In his review of the state of personality assessment for the *Annual Review of Psychology*, Lanyon (1984) offered a stern indictment of the CBTI industry. He noted that available literature regarding these programs appeared to come in three essential varieties. These included:

- (a) glossy promotional literature sometimes masquerading as scientific data and usually accompanied by sample records, (b) studies of customer and user satisfaction, which has never been much of a problem, and (c) an occasional paper giving actual information about the development or validation of an automated system. (p. 690)

Lanyon was particularly distressed by the fact that gross deficits in program validity appeared to have become the norm for CBTI systems. The primary ethical issue of concern here has to do with the manner in which psychologists describe and/or promote computerized assessment services. Two sections of the Ethics Code (APA, 1992) have particular relevance here. First, when describing the nature and results of psychological services, "Psychologists provide, using language that is reasonably understandable to the recipient of those services, appropriate information later about results and conclusions" (p. 1600). Second, the section on avoidance of false or deceptive statements emphasizes that psychologists refrain from making false, deceptive, or misleading statements either by virtue of what they convey or omit concerning their services and work activities. This emphasis on avoiding deception also extends to descriptions of the scientific or clinical basis for, or results or degree of success of psychologists services. The implication of these standards would suggest a rather clear mandate for psychologists to explicitly describe those CBTI services they participate in developing, promoting, or utilizing in their clinical work. This would include descriptions of the process by which the program was constructed, the manner in which it generates interpretive material and any existing evidence of validity. By the same token, psychologists must be proactive in highlighting deficits in system validity or performance such that consumers and users might avoid harmful outcomes. Finally, psychologists should avoid misuse of their influence in the promotion of CBTIs and other computerized assessment

products. "They [Psychologists] are alert to and guard against personal, financial, social, organizational, or political factors that might lead to misuse of their influence" (APA, 1992, p. 1601). Endorsements by respected psychologists in the field of assessment are often coveted and highly promoted by product marketers. The Ethics Code clearly warns against irresponsible promotion of CBTI products in a manner which might compromise the profession or increase misuse of such materials by other professionals and consumers.

4.17.5.4 Competence

Establishing and maintaining an appropriate level of competence in the area of computerized psychological assessment may be one of the most significant areas of ethical risk for psychologists at this time. The Ethics Code (APA, 1992) makes numerous references to the importance of competence for psychologists. The first ethical principle in the Ethics Code relates to competence and stresses that psychologists maintain a high degree of competence in their work and clearly articulate the boundaries of their expertise. "Psychologists function only within boundaries of competence based on education, training and supervised experience or appropriate professional experience" (p. 1599). Further, they provide services in new areas of practice "only after first undertaking appropriate study, training, supervision and/or consultation from persons competent in those areas" (p. 1600). Initially establishing and demonstrating competence in the area of CAPA would not appear to be adequate, however, as psychologists are additionally enjoined by the Ethics Code to maintain their expertise as long as they practice as psychologists. "Psychologists maintain a reasonable level of awareness of current scientific and professional information in their fields of activity and undertake ongoing efforts to maintain competence in the skills they use" (p. 1600).

Development of clear standards and guidelines for use of CBTIs in particular will serve to protect consumers and the profession. They might also serve as critical guides to psychologists, judges, and the courts in determining the standard of practice in this area (Hofer & Green, 1985). Bersoff and Hofer (1991) noted that establishing a prevailing "standard of care" in the area of CBTIs is critical to determining whether a test program's user, developer, or publisher violated a prevailing standard and is therefore ethically out of compliance or legally culpable. Examples of such noncompliance might include negligent entry of data, selection

of a scoring system the psychologist should know is inappropriate for a client or unreasonable reliance on interpretive material from a CBTI narrative report. Problematically, the wide availability of CBTI and other computerized assessment services to persons of varied professional and educational backgrounds (Fowler, 1985), has rendered development of guidelines for competence in this area quite difficult.

The current Ethics Code more explicitly addresses the practice of psychological assessment and offers a clearer picture of how "competence" in this area might be defined:

Those who develop, administer, score, interpret or use psychological assessment instruments do so in a manner and for purposes that are appropriate in light of research or on evidence of the usefulness and proper application of the techniques ... Psychologists who perform interventions or administer, score, interpret, or use assessment techniques are familiar with the reliability, validation, and related standardization or outcome studies of and proper applications and uses of the techniques they use. (APA, 1992, p. 1603)

Although broad, the Ethics Code does suggest several primary areas in which psychologists should have reasonable expertise if they are to competently utilize computer generated assessment material in their work with clients.

As a result of the ethical requirement for practitioners to evaluate the soundness of CBTI reports, they must obviously be qualified to interpret the test themselves. This requires not only basic familiarity with psychometric principles but also a rather detailed understanding of the manner in which the particular system in question was developed and generates interpretive material. On the basis of this understanding, the psychologist might then be able to reject, modify, or expand reports for particular clients (Hofer & Green, 1985). Psychologists will need to be familiar with three components of the CBTI services they utilize in order to do this effectively. These include (i) the examinees score on the relevant test or scale, (ii) the test scale or combination of scales on which interpretations are based, and (iii) research or clinical evidence supporting the interpretation.

In addition to basic competence in psychometrics and interpretation of CBTI assessment findings for individual clients, psychologists must also demonstrate competence in the appropriate monitoring of CBTI data. The Ethics Code requires psychologists to make reasonable efforts to maintain the security of tests and other assessment techniques (APA, 1992). To do this effectively in the area of CAPA, psychologists must establish formal

procedures for retaining, reviewing, and releasing computerized assessment data. Tranel (1994) noted that psychologists bear responsibility for determining whether those requesting test data are qualified to interpret it appropriately. The concern here relates to nonexperts drawing erroneous conclusions based on naive use of CBTI reports. It appears that psychologists must not only avoid irresponsible use of computerized test data themselves, but must also prevent the same on the part of others, "Psychologists do not misuse assessment techniques, interventions, results and interpretations and take reasonable steps to prevent others from misusing the information these techniques provide" (APA, 1992, p. 1603). Related to the release of test data is concern that test questions or CBTI system information may become part of the public domain, resulting in risk of invalidation of tests as well as potential violation of copyright laws and contractual obligations (APA, 1996; Tranel, 1994). Because many CBTI reports include printouts of the client's raw scores as well as those critical items endorsed, users must exercise the same approach to maintaining test security they might employ with any other form of test data.

4.17.5.5 Professional Context

Perhaps the most alluring and potentially dangerous property of CBTI narrative reports is their polished, professional, and thorough appearance. Psychologists, like other users of these services, may be tempted to rely excessively on information from such computerized narratives without adequate interaction with the individual client or reasonable consideration of the unique circumstances in which the client presents for evaluation. The Ethics Code (APA, 1992) rather clearly addresses this concern,

Psychologists perform evaluations, diagnostic services or interventions only within the context of a defined professional relationship ... Psychologists' assessments, recommendations, reports and psychological diagnostic or evaluative statements are based on information and techniques (including personal interviews of the individual when appropriate) sufficient to provide appropriate substantiation for their findings. (p. 1603)

Most concur that computerized assessment narratives must be carefully reviewed for appropriateness and "fit" with the examinee in light of research, complete information about the examinee, and solid professional judgment (Carson, 1990; Fowler, 1985; Graham, 1993; Hofer & Green, 1985).

Matarazzo (1990) made the case that one of the primary sources of ethical and professional danger in this regard has been a rather subtle but progressive loss of distinction between psychological assessment as a professional activity and mere testing.

Objective psychological testing and clinically sanctioned and licensed psychological assessment are vastly different, even though assessment usually includes testing . . . Psychological assessment is engaged in by a clinician and a patient in a one-to-one relationship and has statutorily defined or implied professional responsibilities . . . Specifically, it [assessment] is the activity of a licensed professional, an artisan familiar with the accumulated findings of his or her young science. (p. 1000)

Similarly, Carson (1990) called for defense of “clinicianship” (p. 437) within psychological assessment and highlighted many of the dangers inherent in considering CBTI data apart from other primary client information and without the benefit of a clear client–professional relationship. Those portions of the Ethics Code addressing utilization of assessment results and computerized scoring and interpretive services are clear that psychologists retain full responsibility for conducting competent and context appropriate assessment, versus context blind and potentially harmful psychological testing, “Psychologists retain appropriate responsibility for the appropriate application, interpretation and use of assessment instruments, whether they score and interpret such tests themselves or use automated or other services” (APA, 1992, p. 1604). A related but unresolved concern, however, is how psychologists are to retain such responsibility when most CBTI reports are not signed by a responsible psychologist. Matarazzo (1986) highlighted this problem:

Although it is not yet part of psychology’s code of ethics, my experience leaves no question that computerized clinical interpretations offered in a professional setting about a person’s intellectual, personality, brain-behavior and other highly personal characteristics constitutes a legally and professionally significant invasion of privacy and requires at the least, that the individuals offering these clinical interpretations sign their names to such consultations just as is done in every other profession. (p. 21)

4.17.5.6 CAPA with Special Populations

Related to the foregoing concern regarding the professional context of assessment and the psychologist’s responsibility for ensuring the accuracy and appropriateness of computer

generated assessment material, is concern regarding the implications of CBTIs for special client populations. The APA Ethics Code requires sensitivity to and respect for human differences among examinees and clients. This includes sensitivity to differences across such domains as age, gender, race, ethnicity, national origin, religion, sexual orientation, disability, language, and socioeconomic status. The Ethics Code specifically states that psychologists “remain vigilant for situations in which adjustments must be made in administration or interpretation secondary to individual or contextual factors” (APA, 1992, p. 1603). The rather clear implication here is that computer-assisted interpretive system results must be passed through the clinician’s own interpretive grid with an eye toward identification of demographic or contextual factors on the part of the client which might raise concern about the validity of the findings (Bersoff & Hofer, 1991). This of course demands that the psychologist understands, and has access to, differences in base rates for specific demographic groups with respect to both the test and the interpretive program in question.

As a result of the burgeoning of CBTI systems, the proliferation of unsatisfactory and typically invalidated systems and the widespread marketing of such systems to unqualified users, there have been frequent calls in the scholarly and professional literature for development of standards and regulations relative to computerized approaches to assessment (Burke & Normand, 1987). In 1986 the APA published a set of guidelines for practice by psychologists in this arena. *Guidelines for computer-based test interpretations* (APA, 1986) offered a set of brief and general aspirational guidelines for developers and users of computerized assessment techniques and products. Conscious of the foregoing summary of salient ethical concerns in the field of CAPA, we will now offer a synopsis of those guidelines with a focus on why such guidelines are significant and how they might be applied by psychologists.

4.17.6 GUIDELINES FOR USERS OF COMPUTER-BASED TESTS AND INTERPRETATIONS

The following guidelines are intended for those professionals who use computer-based testing and interpretive services with those to whom they provide services. Table 2 contains the APA Guidelines for users of computer-based tests and interpretations (APA, 1986) and will serve as an outline for the current discussion.

Table 2 Guidelines for users and developers of computer-based tests and interpretations.

<i>Guidelines for users</i>	
<i>Administration</i>	
1.	Influences on test scores due to computer administration that are irrelevant to the purposes of assessment should be eliminated or taken into account in the interpretation of scores.
2.	Any departure from the standard equipment, conditions, or procedures, as described in the test manual or administrative instructions, should be demonstrated not to affect test scores appreciably. Otherwise, appropriate calibration should be undertaken and documented (see Guideline 16).
3.	The environment in which the testing terminal is located should be quiet, comfortable, and free from distractions.
4.	Test items presented on the display screen should be legible and free from noticeable glare.
5.	Equipment should be checked routinely and should be maintained in proper working condition. No test should be administered on faulty equipment. All or part of the test may have to be readministered if the equipment fails while the test is being administered.
6.	Test performance should be monitored, and assistance to the test-taker should be provided, as is needed and appropriate. If technically feasible, the proctor should be signaled automatically when irregularities occur.
7.	Test-takers should be trained on proper use of the computer equipment, and procedures should be established to eliminate any possible effect on test scores due to the test-taker's lack of familiarity with the equipment.
8.	Reasonable accommodations must be made for individuals who may be at an unfair disadvantage in a computer testing situation. In cases where a disadvantage cannot be fully accommodated, scores obtained must be interpreted with appropriate caution.
<i>Interpretation</i>	
9.	Computer-generated interpretive reports should be used only in conjunction with professional judgment. The user should judge for each test-taker the validity of the computerized test report based on the user's professional knowledge of the total context of testing and the test-taker's performance and characteristics.
<i>Guidelines for developers</i>	
<i>Human factors</i>	
10.	Computerized administration normally should provide test-takers with at least the same degree of feedback and editorial control regarding their responses that they would experience in traditional testing formats.
11.	Test-takers should be clearly informed of all performance factors that are relevant to the test result.
12.	The computer testing system should present the test and record responses without causing unnecessary frustration or handicapping the performance of test-takers.
13.	The computer testing system should be designed for easy maintenance and system verification.
14.	The equipment, procedures, and conditions under which the normative, reliability, and validity data were obtained for the computer test should be described clearly enough to permit replication of these conditions.
15.	Appropriate procedures must be established by computerized testing services to ensure the confidentiality of the information and the privacy of the test-taker.
<i>Psychometric properties</i>	
16.	When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when the rank orders of scores of individuals tested in alternative modes closely approximate each other, and the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode.
17.	The validity of the computer version of a test should be established by those developing the test.
18.	Test services should alert test users to the potential problems of nonequivalence when scores on one version of a test are not equivalent to the scores on the version for which norms are provided.
19.	The test developer should report comparison studies of computerized and conventional testing to establish the relative reliability of computerized administration.
20.	The accuracy of computerized scoring and interpretation cannot be assumed. Providers of computerized test services should actively check and control the quality of the hardware and software, including the scoring, algorithms, and other procedures described in the manual.
21.	Computer testing services should provide a manual reporting the rationale and evidence in support of computer-based interpretation of test scores.

Table 2 (continued)**Classification**

22. The classification system used to develop interpretive reports should be sufficiently consistent for its intended purpose (see Chapter 2 of the 1985 *Testing Standards*). For example, in some cases it is important that most test-takers would be placed in the same groups if retested (assuming the behavior in question did not change).
23. Information should be provided to the users of computerized interpretation services concerning the consistency of classifications, including, for example, the number of classifications and the interpretive significance of changes from one classification to adjacent ones.

Validity of computer interpretations

24. The original scores used in developing interpretive statements should be given to test users. The matrix of original responses should be provided or should be available to test users on request, with appropriate considerations for test security and the privacy of test-takers.
25. The manual or, in some cases, interpretive report, should describe how the interpretive statements are derived from the original scores.
26. Interpretive reports should include information about the consistency of interpretations and warnings related to common errors of interpretation.
27. The extent to which statements in an interpretive report are based on quantitative research vs. expert clinical opinion should be delineated.
28. When statements in an interpretive report are based on expert clinical opinion, users should be provided with information that will allow them to weigh the credibility of such opinion.
29. When predictions of particular outcomes or specific recommendations are based on quantitative research, information should be provided showing the empirical relationship between the classification and the probability of criterion behavior in the validation group.
30. Computer testing services should ensure that reports for either users or test-takers are comprehensible and properly delimit the bounds within which accurate conclusions can be drawn by considering variables such as age or sex that moderate interpretations.

Review

31. Adequate information about the system and reasonable access to the system for evaluating responses should be provided to qualified professionals engaged in a scholarly review of the interpretive service. When it is deemed necessary to provide trade secrets, a written agreement of nondisclosure should be made.

4.17.6.1 Administration

Guidelines 1–8 clearly address the responsibility of the computerized system's user in ensuring that examinees are not adversely affected by the computerized administration itself. The environment must be conducive to comfort and maximal performance on the part of the examinee and extraneous influences on his or her performance should be minimized. While computerized administration generally appears to enhance the ease of test taking while decreasing overall time required (Green, 1991b), such programs should routinely assess for evidence that the examinee understands tests and procedures. If an examinee demonstrates any reservation or apprehension regarding interacting with a computer, a conventional format version of the test should be substituted when possible. In the future, a good deal more research is needed to better understand the impact of computers on examinee experience of the assessment process, satisfaction with the experience, and response to the results generated (Finn & Butcher, 1991).

4.17.6.2 Evaluation and Selection of CBTIs

The potential for profound misuse of testing material exists when individuals without adequate training are granted access to automated interpretive systems (Butcher, 1987). When the CBTI user is not a psychologist or lacks the background to effectively and reliably interpret tests, danger exists that CBTIs will not be appropriately scrutinized and carefully selected for the client in question. Moreland, Eyde, Robertson, Primoff, and Most (1995) reported on an attempt to establish basic test user qualifications. The authors found that the 86 identified testing competencies could be distilled to 12 "minimum competencies" and that these could be further reduced to two major categories of competence. Test-users must (i) possess adequate knowledge of the test (or test scoring and interpretation program) and its limitations, and (ii) accept responsibility for the competent use of the test (or computerized test program). Competent use of CBTIs requires that the psychologist carefully evaluate and select appropriate and reasonably validated interpretive programs.

Although well-designed interpretive programs which utilize a careful compilation of empirical data and expert opinion often yield more valid interpretive reports than those generated by clinicians (Green, 1991b), users must become familiar with and conversant regarding the program's established validity. Nonetheless, determining acceptable levels of validity for automated reports poses two problems. First, many tests themselves lack adequate development research and second, most interpretive programs have no established validity themselves or very little (Butcher, 1987). Therefore, it is critical for the prospective user to review the test manual, related documentation, and examples of the computerized reports prior to utilizing them with clients. It is particularly important to evaluate the system rationale for interpreting scores and profiles. How well are the classification and decisional criteria operationalized? Potential users are also encouraged to evaluate the credentials of the system's authors (Lanyon, 1987) and consider the authors' standing as both an expert clinician and a scholar in the field of CBTI.

Another basis on which to evaluate the potential value of a CBTI is that of general utility. As a rule, the rationale or justification for conducting an assessment hinges on provision of information of value with respect to planning and executing treatment (Hayes, Nelson, & Jarrett, 1987). A CBTI will have utility to the extent there is evidence that it contributes to positive treatment outcome, or in some way enhances the status of individuals for whom it is employed. Those systems high in utility will generally be highly efficient and usable, time and cost-effective, and able to discriminate effectively such that real between-examinee differences are detected (Krug, 1987).

Finally, Ben-Porath and Butcher (1986) offered several questions which might be utilized by prospective users of CBTI systems determining which system to employ. We believe these questions offer a good summary of the major concerns expressed in these guidelines. They include (i) to what extent has the validity of the reports been established? (ii) to what extent do the reports rely on empirical findings in generating interpretations? (iii) to what extent do the reports incorporate all of the currently available validated descriptive information? (iv) do the reports take demographic variables into account? (v) are different versions of the report available for different referral questions? (vi) do the reports include practical suggestions? and (vii) are the reports periodically revised to reflect newly acquired information?

4.17.6.3 Interpretation

The APA guidelines (APA, 1986) are clear in warning that CBTIs should never be used as a singular indicator of a person's characteristics, psychological functioning, or diagnosis. Rather, professional judgment is required to determine the extent to which the computerized report is valid and appropriate in light of the total context of testing and test-taker's specific performance and characteristics. Many CBTI program developers and users express similar concern about the danger inherent in divorcing the skilled psychologist from the interpretation and use of computerized reports (Bersoff & Hofer, 1991; Butcher, 1987; Finn & Butcher, 1991; Groth-Marnat & Schumaker, 1989). The polished and objective appearance of narrative reports may lend to the temptation to accept them as valid without adequate scrutiny of their match with the examinee. Bersoff and Hofer (1991) noted

There must be an interposition of human judgment between the CPTI report and decision making to ensure that decisions are made with full sensitivity to the nuances of test administration and interpretation, and that the unique constellation of attributes in each person is evaluated. (p. 241)

Responsible assessment will necessarily involve a multistep process (Finn & Butcher, 1991) in which computerized assessment results are skillfully integrated with other sources of information about the examinee. Currently, computers are not capable of offering a sophisticated synthesis of psychological test results. This integrative function appears to be squarely within the purview and professional responsibility of the psychological diagnostician. Without such integration, psychological reports are necessarily overly general, nonspecific, and questionably accurate. They would certainly fail to capture the test-taker's cognitive, affective, and behavioral functioning across a variety of situations (Bersoff & Hofer, 1991).

The prevailing standard ethically and legally appears to be that the computer-based report is a professional-to-professional consultation (Butcher, 1987). In this way, the computer is merely the equivalent of a library or consultant which might offer the most frequently indicated inferences or correlates for specific test scores or profiles. With the computer as librarian or "look up table" (Butcher, 1987, p. 9), the final determiner of the accuracy and adequacy of the computer-based report is the psychologist who receives the report.

An excellent example of adherence to this guideline is offered by Finn & Butcher (1991)

who quote the disclaimer from the Clinical Interpretive Report for the MMPI. This disclaimer is printed on each interpretive report and highlights for users how the report should be utilized, "This MMPI interpretation can serve as a useful source of hypotheses about clients ... the personality descriptions, inferences, and recommendations contained herein need to be verified by other sources of clinical information since individual clients may not fully match the prototype" (p. 367).

4.17.7 GUIDELINES FOR DEVELOPERS OF COMPUTER-BASED TEST SERVICES

The following guidelines apply most directly to those involved in the construction, validation, and marketing of computerized assessment systems for use by others. Table 2 contains the guidelines for developers of computer-based test services (APA, 1986).

4.17.7.1 Human Factors

Guidelines 10–15 suggest developers of CAPA systems are broadly responsible to ensure that computerized administration of tests does not hamper the examinees' performance in any way. Examinees must also maintain reasonable control of the testing process and the testing service must assume full responsibility for establishing appropriate procedures for ensuring the confidentiality of data collected and the privacy of the examinee. This may be particularly challenging in light of the development of on-line test scoring and interpretation services. On-line access to client information must be carefully controlled by the testing service.

4.17.7.2 Psychometric Properties

Guidelines 16–21 in Table 2 require test or CBTI program developers to carefully evaluate and communicate the psychometric quality of the test or computerized interpretive system. When developing computerized versions of conventional tests, norms and cutting scores for the conventional test can only be used if the computer and conventional form are found to be equivalent (Green, 1991b). Not only should correlations between the two versions be high, but score distributions must be generally equivalent as well. Developers are expected to communicate results of equivalency studies to prospective users and to take initiative in "alerting" users to potential problems resulting from nonequivalence between forms.

4.17.7.3 Classification Strategy

When a computerized assessment system utilizes a classification system based on cutting scores, system developers must offer a convincing rationale for the particular system adopted and the cutting scores selected. In order for a CBTI system to be fully actuarial, system output is determined solely by statistical regularities that have been empirically demonstrated to exist between input and output data (Moreland, 1985). Instead, most systems combine actuarial and clinical expertise approaches. Users must be aware of the system strategy for integrating statistical and clinical prediction in the service of classifying examinees. Additionally, developers bear responsibility for communicating to users the consistency of the classification system and the meaning associated with changes between categories.

4.17.7.4 Validity of Computer Interpretations

As indicated by Guidelines 24–30, developers of computer-based test services are responsible for communicating information to users concerning the system's validity. Not only should validity data be made available to users, but developers must also clarify the extent to which interpretive statements are based on expert clinical opinion or quantitative research. Users must be shown the connection between such research or clinical opinion and specific interpretive statements and classification decisions. Further, developers are to warn users of common errors and potential pitfalls associated with the interpretive system. One of the reasons for the pervasive problem with establishing the validity of CBTI systems has to do with the practice of utilizing clinician generated reports as the criterion in validity research (Moreland, 1985). At times, computer reports may be more accurate than their clinician generated counterparts, thus falsely lowering validity coefficients. Developers should consider alternative criterion variables when possible.

Lanyon (1987) described six factors which should be considered by CBTI system developers in the service of increasing the validity of CBTI programs. First, reliability coefficients for both predictor and criterion variables have been unacceptably low for most CBTIs and should be increased. Second, in balancing the "bandwidth-fidelity" tension, CBTI developers should develop test indices that lead to single, focused predictions (narrow bandwidth/high-fidelity) versus trying to say too much from too little data. Third, departures from empirical data should be minimized. When test interpretations are based on data without attempts

to polish, cluster, or otherwise alter them for the sake of appearance, errors in interpretation are minimized. Fourth, unwarranted generalizations from the standardization sample of CBTIs is a major source of invalidity. When the gap between the population from which an interpretive system was derived and the population on which it is used is substantial, validity declines and erroneous interpretations abound. Fifth, Lanyon recommends that an "unclassifiable" option be vigorously employed. Adding such a category (versus forcing predictive or interpretive statements for every profile) substantially enhances system validity. Finally, whenever possible, different base rates for characteristics being predicted or described should be employed in the interpretive system.

4.17.7.5 Facilitation of Review

The APA guidelines (APA, 1986) highlight the requirement for CBTI system developers to provide adequate information about the system as well as reasonable access to this information on the part of qualified professionals engaged in reviewing the interpretive system. Previous reviewers of this topic appear to concur that availability of detailed development information and data is critical to responsible evaluation of CBTI systems and their usefulness to clinicians (Green, 1991b; Lanyon, 1987; Roid & Gorsuch, 1984; Snyder et al., 1990). High quality interpretive programs clearly label the program using standardized descriptions which clarify the function of the system and the specific manner in which it generates interpretive statements. Such programs provide detailed data relative to development of the system and extensive references to the empirical basis for the decision rules used.

4.17.8 DIRECTIONS FOR THE FUTURE

Various authors have speculated about the future role and developmental course of CAPA. These have included highly optimistic outlooks such as that of Moreland (1985), "I am confident that the computer will eventually replace the professional for most, but not all, assessment functions . . . and this may happen sooner rather than later" (p. 229), as well as more pessimistic perspectives based on fears about the dehumanization of the assessment process (Matarazzo, 1986).

A primary problem in the development in CAPA is the substantial lag in "psychotechnology" currently evident in the assessment field. Specifically, our understanding of assessment appears to lag behind available computer

technology. Along these lines, there is a rather profound need for better research designs and demonstrations of program validity (Lanyon, 1984) among CBTI systems. Currently, a number of such programs offer little in the way of adherence to APA (1986) guidelines for development of CBTI software.

Hopefully, the future holds more promise for empirical research on the reliability and validity of CBTI systems. As more powerful desk-top (and lap-top) computers expand, increasingly complex statistical analyses are possible for test developers who have the expertise to use them. No longer is computer power located in a small number of central or "mainframe" facilities. Greater resources and more creativity in research designs for the study of CBTI are certainly needed. Creative methods of linking test and nontest data, in cost-effective formats, will also be crucial to the development and long-term refinement of CBTI. Test developers should study the classic examples given by Lachar (1984) and the methods reviewed by Snyder et al. (1990) and find ways to connect clinician's ratings, medical and treatment histories, and demographic data to the test-score data file. An important area of expansion in neuropsychological assessment will be the connection of imaging technology (e.g., magnetic resonance imaging [MRI] results) to psychological and cognitive test scores.

Alternatively, the field of CBTI should shift more toward a model of computer as "research assistant" (Roid, 1985), and provide methods for the clinician to access research findings, empirical correlates, and, perhaps, brief and verifiable narrative descriptions such as symptom lists from clinician checklists. However, in the "assistant" model, the final selection and control of narrative statements would be retained as a function of the clinician, not the computer program. As mentioned previously, the word-processing capabilities of systems such as WISC-III Writer (Psychological Corporation, 1994) should make correlated empirical findings available in a "scrapbook" but not be automatically printed in a report. Such systems could become even more elaborate in the future, employing extensive database functions and multimedia graphics, to supplement the conventional test-score results. Much could be done in the future to have even more demographic and historical data available for each client, as long as privacy is protected. The control of the assembly, selection, and composition of the final collection of, for example, data, statements, and graphs, should remain in the hands of the experienced clinician.

Another area of future development will surely be in the expansion of the multimedia

capability of CAPA programs. All types of CAPA, from administration to scoring to interpretation could employ more sophisticated graphic and, perhaps, video capability. Gregory (1996) reports on the development of multimedia "situational" tests being developed at IBM that depict actual on-the-job scenes. The computer may briefly "pause" a video display and ask the examinee to answer questions, or predict the "next step" in the scenario, for purposes of assessing the examinee's sensitivity to interpersonal or technical concerns on the job. For example, the senior author of this chapter assisted in a research project to develop qualification tests for parole officers who were being screened for ability to perform "combative arrests." Film and audio depictions of scenarios with real parole officers were studied to identify critical behaviors such as advanced planning of back-up assistance, prediction of the parolee's potential route in the event of an escape attempt, and physical strategy for placing handcuffs on the subject. Extensive studies of physical movements involved in arrest scenarios were conducted to determine hand and arm strengths required. However, it became clear that many skills were "cognitive" rather than physical and included skills such as planning, prediction, and knowledge of typical parolee strategies of escaping arrest. Such an evaluation might be depicted on computerized video collections, shown to potential applicants, and present questions of strategy and planning. Clearly, extensive empirical validation with actual parole officers, including those who were previously judged to be "expert" in combative arrest, would be essential to the development of such a system. The key element of this "futuristic" system would be the video reenactments, presented on multimedia computer equipment—a method of simulation that has been available on motion-picture film for decades. In the future, more precise computer-adaptive testing may be possible for this medium.

One of the more promising areas for future development is that of adaptive testing. By administering only those test items necessary to draw supportable clinical conclusions, psychologists should be able to substantially reduce testing time while enhancing both the reliability and validity of findings (Krug, 1987). Computer-generated and tailor-made personality tests are a particularly interesting possibility.

Finally, the debate regarding the role of computers in the assessment process will certainly continue in the future around the issue of the role of the clinical psychologist. Butcher (1987) noted

As interest in developing integrated clinical diagnostic reports broadens, more research or system adequacy will be stimulated, and, no doubt, more intense dialogue will be generated on the appropriateness of machines to perform what is believed by some to be an essentially human activity. (p. 11)

4.17.9 CONCLUSIONS AND RECOMMENDATIONS

In summary, the best recommendation for clinical practice, given the current state of validation of narrative computer-interpretive programs is for clinicians to "draw a line" between scoring programs and narrative-interpretive programs. Computer administration of tests and scoring (including elaborate statistical scoring and graphical display) can clearly assist the clinician in terms of efficient use of time and accuracy of calculation. Unless the following conditions are satisfied, it is recommended that CBTI programs with extensive narrative reports be used only for hypothesis generation and never as an unedited section of a psychological report:

(i) *Clinician retains control of word/sentence selection.* If the narrative program gives control to the clinician for assembly of narrative descriptions, then the unique situation of the client can be used to temper the assessment narrative. Automatically printed statements should never be used unedited. Thus, CBTI programs should be "research assistants," not "automated interpreters."

(ii) *Test developers apply actuarial research and document it.* The technical manual accompanying CBTI programs should clearly describe the details of the empirical validation of all descriptive statements, list all cutting scores and their validation studies, present classification accuracy statistics for all profile rules employed, and provide cautionary statements of study limitations.

(iii) *Researchers of CBTI should develop more cost-effective research designs.* Given that the cost of validation is a frequent excuse given for lack of empirical study, statisticians and research experts are encouraged to creatively design new types of studies in cooperation with CBTI developers so the field of CAPA can advance on a more scientific basis where possible.

(iv) *Psychologists should be vigilant in distinguishing assessment from testing.* In the interests of preventing the erosion of the meaning of clinical assessment, as emphasized by Matarazzo (1986, 1990), all psychologists should be alerted to the concern that the final evaluation of all relevant client information, including the context, history, and uniqueness of the client, be reserved for the experienced, trained clinician, not the computer.

4.17.10 REFERENCES

- Adams, K. M., & Shore, D. L. (1976). The accuracy of an automated MMPI interpretation system in a psychiatric setting. *Journal of Clinical Psychology*, 32, 80-82.
- American Psychological Association. (1985). *Standard for educational and psychological testing*. Washington, DC: Author.
- American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington DC: Author.
- American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist*, 47, 1597-1611.
- American Psychological Association. (1996). Statement on the disclosure of test data. *American Psychologist*, 51, 644-648.
- Andrews, L. W., & Gutkin, T. B. (1991). The effects of human versus computer authorship on consumers' perceptions of psychological reports. *Computers in Human Behavior*, 7, 311-317.
- Angoff, W. H., & Huddleston, E. M. (1958). The multi-level experiment: A study of two-stage system for the College Board SAT (Statistical Report No. 58-21). Princeton, NJ: Educational Testing Service.
- Assessment Systems (1990). *MicroCAT testing system manual*. Minneapolis, MN: Author.
- Barclay, J. R. (1983). *Barclay classroom assessment system manual*. Los Angeles: Western Psychological Services.
- Bayroff, A. G., & Seeley, L. C. (1967, June). *An exploratory study of branching tests*. Technical Research Note 188, US Army Behavioral Science Research Laboratory.
- Ben-Porath, Y. S., & Butcher, J. N. (1986). Computers in personality assessment: A brief past, an ebullient present, and an expanding future. *Computers in Human Behavior*, 2, 167-182.
- Bersoff, D. N., & Hofer, P. J. (1991). Legal issues in computerized psychological testing. In T. B. Gutkin & S. L. Wise (Eds.), *The computer and the decision-making process* (pp. 225-244). Hillsdale, NJ: Erlbaum.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- Bloom, B. L. (1992). Computer-assisted psychological intervention: A review and commentary. *Clinical Psychology Review*, 12, 169-197.
- Burke, M. J., & Normand, J. (1987). Computerized psychological testing: Overview and critique. *Professional Psychology: Research and Practice*, 18, 42-51.
- Butcher, J. N. (1985). Introduction to the special series. *Journal of Consulting and Clinical Psychology*, 53, 746-747.
- Butcher, J. N. (1987). The use of computers in psychological assessment: An overview of practices and issues. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 3-14). New York: Basic Books.
- Butcher, J. N., Keller, L. S., & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and Clinical Psychology*, 53, 803-815.
- Carson, R. C. (1990). Assessment: What role the assessor? *Journal of Personality Assessment*, 54, 435-445.
- Conners, K. (1995). *Conners' Continuous Performance Test computer program*. North Tonawanda, NY: Multi Health Systems.
- Cowden, D. J. (1946). An application of sequential sampling to testing students. *Journal of the American Statistical Association*, 41, 547-556.
- Dahlstrom, W. G. (1993). Tests: Small samples, large consequences. *American Psychologist*, 48, 393-399.
- Educational Testing Service (1995). *Learning for tomorrow: 1995 Annual Report*. Princeton, NJ: Author.
- Finn, S. E., & Butcher, J. N. (1991). Clinical objective personality assessment. In M. Hersen, A. E. Kazdin, & A. S. Bellack (Eds.), *The Clinical Psychology Handbook* (pp. 362-373). New York: Pergamon.
- Flanagan, J. C., & Lindquist, E. F. (Eds.) (1951). *Educational measurement*. Washington, DC: American Council of Education.
- Fowler, R. D. (1985). Landmarks in computer-assisted psychological assessment. *Journal of Consulting and Clinical Psychology*, 53, 748-759.
- Fowler, R. D., & Butcher, J. N. (1986). Critique of Matarazzo's view on computerized psychological testing. *American Psychologist*, 41, 94-96.
- Garb, H. N. (1994). Judgment research: Implications for clinical practice and testimony in court. *Applied and Preventive Psychology*, 3, 173-183.
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, 23, 483-496.
- Graham, J. R. (1993). *MMPI-2: Assessing personality and psychopathology*. New York: Oxford University Press.
- Green, B. F. (1991a). Computer based adaptive testing in 1991. *Psychology & Marketing*, 8(4), 243-257.
- Green, B. F. (1991b). Guidelines for computer testing. In T. B. Gutkin & S. L. Wise (Eds.), *The computer and the decision-making process* (pp. 245-274). Hillsdale, NJ: Erlbaum.
- Gregory, R. J. (1996). *Psychological testing: History, principles, and applications*. Boston: Allyn and Bacon.
- Groth-Marnat, G., & Schumaker, J. (1989). Computer-based psychological testing: Issues and guidelines. *American Journal of Orthopsychiatry*, 59, 257-263.
- Guastello, S. J., & Rieke, M. L. (1994). Computer-based test interpretations as expert systems. *Computers in Human Behavior*, 10, 435-455.
- Guastello, S. J., Guastello, D. D., & Craft, L. L. (1989). Assessment of the Barnum effect in computer-based test interpretations. *The Journal of Psychology*, 123, 477-484.
- Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment. *American Psychologist*, 42, 963-974.
- Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology*, 53, 826-838.
- Huba, G. J. (1986). Interval banded profile analysis: A method for matching score profiles to "soft" prototypic patterns. *Educational and Psychological Measurement*, 46, 565-570.
- Jackson, D. N. (1985). Computer-based personality testing. *Computers in Human Behavior*, 1, 255-264.
- Kaufman, A. S. (1994). *Intelligent testing with the WISC-III*. New York: Wiley.
- Kaufman, A. S. (1996). *Wechsler integrated interpretive system*. Odessa, FL: Psychological Assessment Resources.
- Klinger, D. E., Miller, D., Johnson, J., & Williams, T. (1977). Process evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. *Behavior Research Methods and Instrumentation*, 9, 110-116.
- Kleinmuntz, B. (1969). Personality test interpretation by computer and clinician. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications* (pp. 97-104). New York: McGraw-Hill.
- Kleinmuntz, B. (1975). The computer as clinician. *American Psychologist*, 30, 379-387.
- Krug, S. E. (1984). *Psychware: A reference guide to computer-based products*. Kansas City, MO: Test Corporation of America.
- Krug, S. E. (1987). Microtrends: An orientation to

- computerized assessment. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 15–25). New York: Basic Books.
- Lachar, D. (1984). *WPS Test Report for the Personality Inventory for Children*. Los Angeles: Western Psychological Services.
- Lanyon, R. I. (1984). Personality assessment. *Annual Review of Psychology*, 35, 667–701.
- Lanyon, R. I. (1987). The validity of computer-based personality assessment products: Recommendations for the future. *Computers in Human Behavior*, 3, 225–238.
- Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, 9, 563–564.
- Linn, R. L., Rock, D., & Cleary, A. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 29, 129–146.
- Lord, F. M. (1968). Some test theory for tailored testing. *Research Bulletin RB-68-38*. Princeton, NJ: Educational Testing Service.
- Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement*, 8, 147–151.
- Marks, P. A., & Seeman, W. (1963). The actuarial description of abnormal personality. Baltimore: Williams and Wilkins.
- Matarazzo, J. D. (1983). Computerized psychological testing (Editorial). *Science*, 221, 323.
- Matarazzo, J. D. (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist*, 41, 14–24.
- Matarazzo, J. D. (1990). Psychological assessment versus psychological testing. *American Psychologist*, 45, 999–1017.
- McBride, J. R. (1988, August). *A computerized adaptive version of the Differential Aptitude Test*. Paper presented at the annual meeting of the American Psychological Association, Atlanta.
- Meehl, P. E. (1954). *Clinical vs. statistical prediction*. Minneapolis, MN: University of Minnesota Press.
- Meehl, P. E. (1956). Wanted: A good cookbook. *American Psychologist*, 11, 263–272.
- Mitchell, J. V., & Kramer, J. J. (1985). Computer-based assessment and the public interest: An examination of the issues and introduction to the special issue. *Computers in Human Behavior*, 1, 203–305.
- Moreland, K. L. (1985). Computer-assisted psychological assessment in 1986: A practical guide. *Computers in Human Behavior*, 1, 221–233.
- Moreland, K. L. (1992). Computer-assisted psychological assessment. In M. Zeidner and R. Most (Eds.) *Psychological testing: An inside view*. Palo Alto, CA: Consulting Psychologists Press.
- Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S., & Most, R. B. (1995). Assessment of test user qualifications. *American Psychologist*, 50, 14–23.
- O'Dell, J. W. (1972). P. T. Barnum explores the computer. *Journal of Consulting and Clinical Psychology*, 38, 270–273.
- Patterson, J. J. (1962). *An evaluation of the sequential method of psychological testing*. Unpublished doctoral dissertation, Michigan State University.
- Piotrowski, C., & Keller, J. W. (1989). Use of assessment in mental health clinics and services. *Psychological Reports*, 64, 1298.
- Pohl, N. F. (1981). Scale considerations in using vague quantifiers. *Journal of Experimental Education*, 49, 235–240.
- Powell, D., Kaplan, E., Whitla, D., Weintraub, S., Catlin, R., & Funkenstein, H. (1993). *MicroCog: Assessment of cognitive functioning*. San Antonio, TX: Psychological Corporation.
- Prince, R. J., & Guastello, S. J. (1990). The Barnum effect in a computerized Rorschach interpretation system. *The Journal of Psychology*, 124, 217–222.
- Psychological Corporation (1987). *Differential aptitude tests computerized adaptive edition manual*. San Antonio, TX: Author.
- Psychological Corporation (1992). *Weschler scoring assistant manual*. San Antonio, TX: Author.
- Psychological Corporation (1994). *WISC-III Writer manual*. San Antonio, TX: Author.
- Rasch, G. (1980). *Some probability models for aptitude and attainment tests*. Chicago: University of Chicago Press.
- Roid, G. H. (1969). *Branching methods for constructing psychological test scales*. Unpublished doctoral dissertation, University of Oregon.
- Roid, G. H. (1985). Computer-based test interpretation: The potential of quantitative methods of test interpretation. *Computers in Human Behavior*, 1, 207–219.
- Roid, G. H. (1986). Computer technology in testing. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 29–69). Hillsdale, NJ: Erlbaum.
- Roid, G. H., & Fitts, W. H. (1988). *Tennessee Self Concept Scale revised manual*. Los Angeles: Western Psychological Services.
- Roid, G. H., & Gorsuch, R. L. (1984). Development and clinical use of test-interpretive programs on microcomputers. In M. D. Schwartz (Ed.), *Using computers in clinical practice* (pp. 141–149). New York: Haworth Press.
- Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1991). Comparability of computerized adaptive and conventional testing with the MMPI-2. *Journal of Personality Assessment*, 57, 278–290.
- Sands, W. A., & Gade, P. A. (1983). An application of computerized adaptive testing in Army recruiting. *Journal of Computer-Based Instruction*, 10, 37–89.
- Silverstein, A. B. (1981). Reliability and abnormality of test score differences. *Journal of Clinical Psychology*, 37, 392–394.
- Snyder, D. (1981). *Manual for the marital satisfaction inventory*. Los Angeles: Western Psychological Services.
- Snyder, D., Widiger, T., & Hoover, D. (1990). Methodological considerations in validating computer-based test interpretations: Controlling for response bias. *Psychological Assessment*, 2, 470–477.
- Sympton, J. B., Weiss, D. J., & Ree, M. J. (1982). *Predictive validity of conventional and adaptive tests in an Air Force training environment (AFHRL TR 81-40)*. Brook Air Force Base, TX: Manpower and Personnel Division, Air Force Human Relations Laboratory.
- Tranel, D. (1994). The release of psychological data to non experts: Ethical and legal considerations. *Professional Psychology: Research and Practice*, 25, 33–38.
- Wainer, H. (Ed.) (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.
- Wald, A. (1947). *Sequential analysis*. New York: Wiley.
- Weiss, D. J. (Ed.) (1983). *New horizons in testing*. New York: Academic Press.
- Wise, S. L., & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and evaluation in counseling and development*, 23, 3–9.

4.18

Therapeutic Assessment: Linking Assessment and Treatment

MARK E. MARUISH

Strategic Advantage, Minneapolis, MN, USA

4.18.1 INTRODUCTION	526
4.18.2 THE CURRENT PRACTICE OF PSYCHOLOGICAL ASSESSMENT IN THE THERAPEUTIC ENVIRONMENT	528
4.18.3 PSYCHOLOGICAL ASSESSMENT AS A THERAPEUTIC ADJUNCT	530
4.18.3.1 <i>Psychological Assessment for Clinical Decision-making</i>	530
4.18.3.2 <i>Psychological Assessment as a Treatment Technique</i>	530
4.18.3.3 <i>Psychological Assessment for Outcomes Assessment</i>	531
4.18.4 GENERAL CONSIDERATIONS FOR THE SELECTION AND USE OF PSYCHOLOGICAL TEST INSTRUMENTATION	531
4.18.4.1 <i>Types of Instrumentation for Therapeutic Assessment</i>	532
4.18.4.1.1 <i>Psychological/psychiatric symptom measures</i>	532
4.18.4.1.2 <i>Measures of general health status and role functioning</i>	533
4.18.4.1.3 <i>Quality of life measures</i>	534
4.18.4.1.4 <i>Service satisfaction measures</i>	534
4.18.4.2 <i>Guidelines for Instrument Selection</i>	535
4.18.4.2.1 <i>National Institute of Mental Health criteria</i>	535
4.18.4.2.2 <i>Other criteria and considerations</i>	536
4.18.5 PSYCHOLOGICAL ASSESSMENT AS A TOOL FOR SCREENING	538
4.18.5.1 <i>Research-based Use of Psychological Screeners</i>	539
4.18.5.2 <i>Implementation of Screeners into the Daily Work Flow of Service Delivery</i>	540
4.18.6 PSYCHOLOGICAL ASSESSMENT AS A TOOL FOR TREATMENT PLANNING	541
4.18.6.1 <i>Assumptions About Treatment Planning</i>	541
4.18.6.2 <i>The Benefits of Psychological Assessment for Treatment Planning</i>	542
4.18.6.2.1 <i>Problem identification</i>	542
4.18.6.2.2 <i>Problem clarification</i>	542
4.18.6.2.3 <i>Identification of important patient characteristics</i>	543
4.18.6.2.4 <i>Monitoring of progress along the path of expected improvement</i>	544
4.18.7 PSYCHOLOGICAL ASSESSMENT AS A THERAPEUTIC INTERVENTION	545
4.18.7.1 <i>What Is Therapeutic Assessment?</i>	545
4.18.7.2 <i>The Impetus for Therapeutic Assessment</i>	546
4.18.7.3 <i>The Therapeutic Assessment Process</i>	546
4.18.7.3.1 <i>Step 1: The initial interview</i>	547
4.18.7.3.2 <i>Step 2: Preparing for the feedback session</i>	547
4.18.7.3.3 <i>Step 3: The feedback session</i>	547
4.18.7.3.4 <i>Additional steps</i>	548
4.18.7.4 <i>Empirical Support for Therapeutic Assessment</i>	548
4.18.8 PSYCHOLOGICAL ASSESSMENT AS A TOOL FOR OUTCOMES MANAGEMENT	549
4.18.8.1 <i>What Are Outcomes?</i>	549
4.18.8.2 <i>Outcomes Assessment: Measurement, Monitoring, and Management</i>	550

4.18.8.3	<i>The Benefits of Outcomes Assessment</i>	550
4.18.8.4	<i>The Therapeutic Use of Outcomes Assessment</i>	550
4.18.8.4.1	<i>Purpose of the outcomes assessment</i>	551
4.18.8.4.2	<i>What to measure</i>	551
4.18.8.4.3	<i>How to measure</i>	552
4.18.8.4.4	<i>When to measure</i>	553
4.18.8.4.5	<i>How to analyze outcomes data</i>	554
4.18.9	FUTURE DIRECTIONS	555
4.18.9.1	<i>What the Industry Is Moving Away From?</i>	555
4.18.9.2	<i>Trends in Instrumentation</i>	556
4.18.9.3	<i>Trends in Data Use and Storage</i>	556
4.18.9.4	<i>Trends in the Application of Technology</i>	557
4.18.10	SUMMARY	558
4.18.11	REFERENCES	559

4.18.1 INTRODUCTION

The cost of health care in the USA has reached astronomical heights. In 1995, approximately \$1 trillion, or 14.9% of the gross domestic product, was spent on health care, and a 20% increase is expected by the year 2000 (*Mental Health Weekly*, 1996a). The cost and prevalence of mental health problems and the accompanying need for behavioral health care services in the USA continue to rise at rates which give cause for concern. America's mental health bill in 1990 was \$147 billion (*Mental Health Weekly*, 1996c). The Center for Disease Control and Prevention (1994) recently reported on the results of a survey of 45 000 randomly interviewed Americans regarding their quality of life. The survey found that one-third of the respondents reported they suffered from depression, stress, or emotional problems at least one day a month, and 11% percent of the sample reported having these problems more than eight days a month.

The American Psychological Association (APA; 1996) also reports statistics, summarized below, that bear attention.

(i) It is estimated that 15–18% of Americans suffer from a mental disorder; 14 million of these individuals are children.

(ii) Approximately eight million Americans suffer from depression in any given one-month period.

(iii) As many as 20% of Americans will suffer one or more major episodes of depression during their lifetime.

(iv) An estimated 80% of elderly residents in Medicaid facilities were found to have moderate to intensive needs for mental health services. Moreover, information from various studies indicates that at least 25% of primary health care patients have a diagnosable behavioral disorder (*Mental Health Weekly*, 1996b).

The need for behavioral health care services is significant. In analyzing data from a 1987

national survey of 40 000 people in 16 000 households, Olfson and Pincus (1994a, 1994b) found that 3% of the population was seen for at least one psychotherapeutic session that year. Of these visits, 81% were to mental health professionals. Estimates provided by Vanden-Bos, DeLeon, and Belar (1993) in the early 1990s indicated that in any year, 37.5 million Americans (or 15% of the population at that time) could benefit from mental health services.

What is the value of the services provided to those suffering from mental illness or substance abuse/addiction/dependency? Some might argue that the benefit is either minimal, or too costly to achieve if significant effects are to be gained. This is in the face of data which suggest otherwise. Numerous studies have demonstrated that treatment of mental health and substance abuse/dependency problems can result in substantial savings when viewed from a number of perspectives. This "cost offset" effect probably has been demonstrated most clearly in savings in medical care dollars over given periods of time.

Medical cost offset considerations are significant, given reports that 50–70% of usual primary care visits are for medical problems that involve psychological factors (APA, 1996). APA also reports that 25% of patients seen by primary care physicians have a disabling psychological disorder, and that depression and anxiety rank among the top six conditions seen by family physicians. Following are just a few of the findings supporting the medical cost benefits that can accrue from providing behavioral health care treatment.

(i) At least 25% or more of patients seen in a primary care setting have diagnosable behavioral disorders and use two to four times as many medical resources as those patients without these disorders (*Mental Health Weekly*, 1996b).

(ii) Sipkoff (1995) reported several conclusions, drawn from a review of numerous studies

conducted between 1988 and 1994 and listed in the *Cost of addictive and mental disorders and effectiveness of treatment* report published by the Substance Abuse and Mental Health Services Administration (SAMHSA). One conclusion derived from a meta-analysis of offset effect was that treatment for mental health problems results in an approximately 20% reduction in the overall cost of health care. The report also concluded that while alcoholics were found to spend twice as much on health care as those without abuse problems, one-half of the cost of substance abuse treatment is offset within one year by subsequent reductions in the combined medical cost savings for the patient and his or her family.

(iii) Strain et al. (1991) found that screening a group of 452 elderly hip fracture patients for psychiatric disorders prior to surgery and providing mental health treatment to the 60% of the sample needing treatment reduced total medical expenses by \$270 000. The cost of the psychological/psychiatric services provided to this group was only \$40 000.

(iv) Simmons, Avant, Demski, and Parisher (1988) compared the average medical costs for chronic back pain patients at a multidimensional pain center (providing psychological and other types of intervention) during the year prior to treatment to those costs of the year following treatment. The pretreatment costs per patient were \$13 284 while post-treatment costs were \$5596.

The reader is referred to Friedman, Sobel, Myers, Caudill, and Benson (1995) for a detailed discussion of various ways in which behavioral interventions can both maximize care to medical patients and achieve significant economic gains.

APA (1996) has very succinctly summarized what appears to be the prevalent findings of the medical cost offset literature.

(i) Patients with mental disorders are heavy users of medical services, averaging twice as many visits to their primary care physicians as patients without mental disorders.

(ii) When appropriate mental health services are made available, this heavy use of the system often decreases, resulting in overall health savings.

(iii) Cost offset studies show a decrease in total health care costs following mental health interventions even when the cost of the intervention is included.

(iv) In addition, cost offset increases over time, largely because ... patients continue to decrease their overall use of the health care system, and don't require additional mental health services. (p. 2)

Medical cost offset effects are relatively obvious and easy to measure. Benefits, financial

and otherwise, that accrue from the treatment of mental health and substance abuse/dependency problems also can come in forms that may not be so obvious. One area in which treatment can have a tremendous impact is that of the workplace. For example, note a few of the facts assembled by APA (1996).

(i) In 1985 behavioral health problems resulted in over \$77 billion in lost income to Americans.

(ii) California's stress-related disability claims totaled \$350 million in 1989.

(iii) In 1980, alcoholism resulted in over 500 million lost work days in this country.

(iv) Major depression cost an estimated \$23 billion in lost work days in 1990. In addition, individuals with this disorder are three times more likely than nondepressed individuals to miss time from work and four times more likely to take disability days.

(v) Of all subjects from 58 psychotherapy effectiveness studies focusing on the treatment of depression, 77% received significantly better work evaluations than depressed subjects who did not receive treatment.

(vi) Treatment resulted in a 150% increase in earned income for alcoholics and a 390% increase in income for drug abusers in one study of 742 substance abusers.

In related findings, anxiety disorders accounted for one-third of America's \$147 billion mental health bill in 1990 (*Mental Health Weekly*, 1996c). And on another front, the former director of the Office of the National Drug Control Policy reported that for every dollar spent on drug treatment, America saves seven dollars in health care and criminal justice costs (*Substance Abuse Funding News*, 1995).

Society's need for behavioral health care services provides an opportunity for trained providers of mental health services to become part of the solution to a major health care problem that shows no indication of decline. Each of the helping professions has the potential to make a particular contribution to this solution. Not the least of these contributions are those that can be made by clinical psychologists. As pointed out in an earlier volume (Maruish, 1994), the use of psychological tests in the assessment of the human condition is one of the hallmarks of clinical psychology. In fact, the training and acquired level of expertise in psychological testing distinguishes the clinical psychologist from other behavioral health care professionals probably more than anything else. Indeed, expertise in test-based psychological assessment can be said to be the particular and unique contribution that clinical psychologists make to the behavioral health care field.

For decades, clinical psychologists and other behavioral health care providers have come to rely on psychological assessment as a standard tool to be used with other sources of information for diagnostic and treatment planning purposes. However, changes that have taken place in the delivery of health care in general, and behavioral health care services in particular, during the past several years have led to changes in the way in which third-party payers and clinical psychologists themselves think about and/or use psychological assessment in day-to-day clinical practice. Some question the value of psychological assessment in the current time-limited, capitated service delivery arena where the focus has changed from clinical priorities to fiscal priorities (Sederer, Dickey, & Hermann, 1996). Others argue that it is in just such an arena that the benefits of psychological assessment can be most fully realized and contribute significantly to the delivery of cost-effective treatment for behavioral health disorders. Consequently, it could assist the health care industry in appropriately controlling or possibly reducing the utilization and cost of health care over the long term. It is this latter side of the argument that is supported by this author, and it provides the basis for this chapter.

In developing this chapter, the intent has been to provide students and practitioners of clinical psychology with an overview of how psychological assessment could and should be used in this era of managed behavioral health care. In doing so, this author discusses how psychological assessment is currently being used in the therapeutic environment and the many ways in which it might be used to the ultimate benefit of patients, providers, and payers.

As a final introductory note, it is important for the reader to understand that the term “psychological assessment,” as it is used in this chapter, refers to the evaluation of a patient’s mental health status using psychological tests or related instrumentation. This evaluation may be conducted with or without the benefit of patient or collateral interviews, review of medical or other records, and/or other sources of relevant information about the patient.

4.18.2 THE CURRENT PRACTICE OF PSYCHOLOGICAL ASSESSMENT IN THE THERAPEUTIC ENVIRONMENT

For a number of decades, psychological assessment has been viewed as a valued and integral part of the services offered by clinical psychologists. However, its popularity has not been without its ups and downs. Megargee and

Spielberger (1992) have described a decrease in interest in assessment that began in the 1960s. This was due to a number of factors, including shifts in focus to those aspects of treatment for which assessment was thought to contribute little. Examples of these aspects included a growing emphasis on behavior modification techniques, the increasing use of psychotropic medications, and an emphasis in studying symptoms rather than personality syndromes and structures. Fortunately, Megargee and Spielberger also noted a number of factors that indicate a relatively recent resurgence in the interest in assessment, including a new realization of how psychological assessment can assist in interventions provided to mental health care patients.

But where does psychological assessment actually fit into the daily scope of activities for practicing psychologists? The results of two recent surveys provide inconsistent findings. The newsletter *Psychotherapy Finances* (1995) reported the results of a nationwide readership survey of 1700 mental health providers of various professions. In this survey, 67% of the participating psychologists reported that they provide psychological testing services. This represents about a 10% drop from a similar survey published in 1992 by the same publication. Also of interest in this survey is the percent of professional counselors (39%), marriage and family counselors (16%), psychiatrists (21%), and social workers (13%) offering these same services.

In a 1995 survey conducted by the APA’s Committee for the Advancement of Professional Practice (Phelps, 1996), 14,000 practitioners responded to questions related to workplace settings, areas of practice concerns, and range of activities. Most of the respondents (40.7%) were practitioners whose primary work setting was an individual independent practice. Other general work settings, that is, government, medical, academic, group practice settings, were represented by fairly equal numbers of respondents from the remainder of the sample. The principal professional activity reported by the respondents was psychotherapy, with 43.9% of the sample acknowledging involvement in this service. Assessment was the second most prevalent activity, being reported by 14% of the sample.

Differences in the two samples utilized in the above surveys may account for the inconsistencies in their findings. Psychologists who are subscribers to *Psychotherapy Finances* may represent that subsample of the APA survey respondents who are more involved in the delivery of clinical services. Certainly the fact that only about 44% of the APA respondents

offer psychotherapy services supports this hypothesis.

Regardless of the two sets of findings, psychological assessment does not appear to be utilized as much as in the past, and one does not have to look hard to determine at least one reason why. One of the major changes that has come about in the American health care system during the past several years has been the creation and proliferation of managed care organizations (MCOs). The most significant direct effects of managed care include reductions in the length and amount of service, reductions in accessibility to particular modalities (e.g., reduced number of outpatient visits per case), and profession-related changes in the types of services managed by behavioral health care providers (Oss, 1996). Overall, the impact of managed behavioral health care on the services offered by psychologists and other health care providers has been tremendous. In the APA survey reported above (Phelps, 1996), approximately 79% of the respondents reported that managed care had either a low, medium, or high negative impact on their work. How has managed care negatively impacted the use of psychological assessment? It is not clear from the results of this survey, but perhaps others can offer at least a partial explanation.

Ficken (1995) has provided some insight into how the advent of managed care has limited the reimbursement for (and therefore the use of) psychological assessment. In general, he sees the primary reason for this as being a financial one. In an era of capitated behavioral health care coverage, the amount of money available for behavioral health care treatment is limited. MCOs therefore require a demonstration that the amount of money spent for testing will result in a greater amount of treatment cost savings. In addition, Ficken notes that much of the information obtained from psychological assessment is not relevant to the treatment of patients within an MCO environment. Understandably, MCOs are reluctant to pay for the gathering of such information.

Werthman (1995) provides similar insights into this issue, noting that

Managed care ... has caused [psychologists] to revisit the medical necessity and efficacy of their testing practices. Currently, the emphasis is on the use of highly targeted and focused psychological and neuropsychological testing to sharply define the "problems" to be treated, the degree of impairment, the level of care to be provided and the treatment plan to be implemented.

The high specificity and "problem-solving" approach of such testing reflects MCOs' commitment to effecting therapeutic change, as opposed

to obtaining a descriptive narrative with scores. In this context, testing is perceived as a strong tool for assisting the primary provider in more accurately determining patient "impairments" and how to "repair" them. (p. 15)

In general, Werthman views psychological assessment as being no different from other forms of patient care, thus making it subject to the same scrutiny, demands for demonstrating medical necessity and/or utility, and consequent limitations imposed by MCOs on other covered services.

The foregoing representations of the current state of psychological assessment in behavioral health care delivery could be viewed as an omen of worse things to come. In this author's opinion, they are not. Rather, the limitations that are being imposed on psychological assessment and the demand for justification of its use in clinical practice represent part of the customers' dissatisfaction with the way things always have been done in the past. In general, this author views the tightening of the purse strings as a positive move for both behavioral health care and the profession of psychology. It is a wake-up call to those who have contributed to the health care crisis by either uncritically performing costly psychological assessments, being unaccountable to the payers and recipients of our services, and generally not performing our services in the most responsible, cost-effective and efficient way possible. It is telling us that we need to evaluate what we've done and the way we've done it, and to determine what is the best way to do it in the future. As such, it provides an opportunity for clinical psychologists to re-establish the valuable contributions they can make to improving the quality of behavioral health care delivery through their knowledge and skills in the area of psychological assessment.

In the sections that follow, this author will convey what he sees are the opportunities for psychological assessment in the behavioral health care arena, both in the present and the future, and the means of best achieving them. The views that are advanced are based on his knowledge of and experience in current psychological assessment practices as well as directions provided by the current literature. Some will probably disagree with the proposed approach, given their own experience and thinking on the matters discussed. However, it is hoped that even though in disagreement, the reader will be challenged to defend his or her position to themselves and as a result, feel more comfortable in their thinking about their approach to their psychological assessment practices.

4.18.3 PSYCHOLOGICAL ASSESSMENT AS A THERAPEUTIC ADJUNCT

The role of psychological assessment in the therapeutic environment traditionally has been quite limited. Those of us who did not receive our graduate clinical training within the past few years probably have been taught the value of psychological assessment only at the “front end” of treatment. We were instructed in the power and utility of psychological assessment as a means of assisting in the identification of symptoms and their severity, personality characteristics relevant to understanding the patient and his or her typical way of perceiving and interacting with the world, and other aspects of the individual (e.g., intelligence, vocational interests) that are important in arriving at a description of the patient at one particular point in time. Based on these data and information obtained from patient and collateral interviews, medical records and the individual’s stated goals for treatment, a diagnostic impression was given and a treatment plan was probably formulated and placed in the patient’s chart, hopefully to be reviewed at various points during the course of treatment. In some cases, the patient was assigned to another practitioner within the same organization or referred out, never to be contacted or seen again, much less be assessed again by the one who performed the original assessment.

Fortunately, during the past few years the usefulness of psychological assessment as more than just a tool to be used at the beginning of treatment has come to be recognized. Consequently, its utility has been extended beyond being a mere tool for describing an individual presenting themselves for treatment to being a means of facilitating the treatment and understanding of behavioral health care problems throughout the episode of care and beyond. Psychologists and others who employ it in their practices are now finding that psychological assessment can be used for a variety of purposes. Generally speaking, several psychological tests currently being marketed can be employed as tools for assisting in clinical decision-making, outcomes assessment and, more directly, as treatment techniques in and of themselves. Each of these uses can uniquely contribute incremental value to the therapeutic process.

4.18.3.1 Psychological Assessment for Clinical Decision-making

Traditionally, psychological assessment has been used to assist clinical psychologists and other behavioral health care clinicians in making important clinical decisions. The types

of decision-making for which it has been used include those related to screening, treatment planning, and monitoring of treatment progress. Generally, screening may be undertaken to assist in either: (i) identifying the patient’s need for a particular service, or (ii) determining the likelihood of the presence of a particular disorder or other behavioral/emotional/psychological problem. More often than not, a positive finding on screening leads to a more extensive evaluation of the patient in order to confirm with greater certainty the existence of the problem, or to further delineate the problem. The value of screening lies in the fact that it permits the clinicians to identify, quickly and economically, with a fairly high degree of confidence (depending on the particular instrumentation used), those who are and are not likely to need care or at least further evaluation.

In many instances, psychological assessment is performed in order to obtain information that is deemed useful in the development of a specific plan for treatment. Typically, it is the type of information that is not easily (if at all) accessible through other means or sources. It is information which, when combined with other information about the patient, aids in understanding the patient, identifying the most important problems and issues that need to be addressed, and formulating recommendations about the best means of addressing them.

Another way in which psychological assessment can play a role in clinical decision-making is in the area of treatment monitoring. Repeated assessment of the patient at regular intervals during the treatment process can provide the therapist with feedback regarding the progress which is being made in the therapeutic endeavor. Based on the findings, the therapist will be encouraged either to continue with the original therapeutic approach or, in the case of no change or exacerbation of the problem, to modify or abandon the approach in favor of an alternate one.

4.18.3.2 Psychological Assessment as a Treatment Technique

It is only recently that empirical studies and other articles addressing the therapeutic benefits that can be realized directly from discussing psychological assessment results with the patient have been published. Rather than just providing test feedback as directed by APA’s *Ethical principles of psychologists* (APA, 1992), therapeutic use of assessment involves a presentation of assessment results (including assessment materials such as test protocols, profile forms, other assessment summary materials) directly to the patient; an elicitation of

the patient's reactions to them; and an in-depth discussion of the meaning of the results in terms of patient-defined assessment goals. In essence, the assessment data can serve as a catalyst for the therapeutic encounter via the objective feedback that is provided to the patient, the patient self-assessment that is stimulated, and the opportunity for patient and therapist to arrive at mutually agreed upon therapeutic goals, based on impressionistic and objective data available to both parties.

4.18.3.3 Psychological Assessment for Outcomes Assessment

Currently, one of the most common reasons for conducting psychological assessment in the USA is to assess the outcomes of behavioral health care treatment. It is difficult to open a trade paper or health care newsletter or to attend a professional conference without being presented with a discussion on either how to "do outcomes" or what the results of a certain facility's outcomes study have revealed. The focus on outcomes assessment most probably can be traced to the "continuous quality improvement" (CQI) movement that was initially implemented in business and industrial settings. The impetus for the movement originally was a desire to produce quality products in the most efficient manner, resulting in increased revenues and decreased costs.

In the health care arena, outcomes assessment has multiple purposes, not the least of which is as a tool for marketing the organization's services. Related to this, those organizations vying for lucrative contracts from third-party payers to provide health care services to their covered lives frequently require outcomes data demonstrating the effectiveness of the services offered by the bidders. Equally important to those awarding contracts is how satisfied patients are with the provider's services. But probably the most important potential use of this data for provider organizations (although not always recognized as such) can be found in the knowledge it yields about what works and what doesn't. In this regard it can serve a program evaluation function. It is this knowledge that, if attended to and acted upon, can lead to improvement in the services the organization offers. When used in this manner, outcomes assessment can become an integral component of the organization's CQI initiative.

But more importantly for the individual patient, outcomes assessment provides a means of objectively measuring how much improvement he or she has made from the time of treatment initiation to the time of treatment termination. Feedback to this effect may serve

to instill in the patient greater self-confidence and self-esteem, and/or a more realistic view of where he or she is (from a psychological standpoint) at that particular time in their life. Conversely, it may serve as an objective indicator to the patient of the need for continued treatment.

The purpose of the foregoing is to present a broad overview of psychological assessment as a multipurpose behavioral health care tool. Depending on the individual clinician or provider organization, it may be employed for one or more, or all, of the purposes just described. Knowing the various ways in which psychological assessment can be used in the service of therapeutic change should help the reader understand the more in-depth and detailed discussion about how these applications can facilitate or otherwise add value to the psychotherapeutic services offered by providers. This detailed discussion follows below.

Before beginning this discussion, however, it is important to briefly review the types of instrumentation most likely to be used in therapeutic psychological assessment, as well as the significant considerations and issues related to the selection and use of this instrumentation for the stated purposes. This should further facilitate the reader's understanding of the remainder of the chapter.

4.18.4 GENERAL CONSIDERATIONS FOR THE SELECTION AND USE OF PSYCHOLOGICAL TEST INSTRUMENTATION

New instrumentation for facilitating and evaluating behavioral health care treatment is released by the major test publishers annually. Thus, the availability of instrumentation for these purposes is not an issue. However, selection of the appropriate instrument(s) for one or more of the therapeutic purposes described above is a matter requiring careful consideration. Inattention to the instrument's intended use, its demonstrated psychometric characteristics, its limitations, and other aspects related to its practical application can result in misguided treatment and potentially harmful consequences for a patient.

Several types of instruments could be used for the general therapeutic assessment purposes described above. For example, neuropsychological instruments might be used to assess memory deficits that could impact the clinician's decision to perform further testing, the goals established for treatment, and the approach to treatment that is selected. Tests designed to provide estimates of level of

intelligence might be used for the same purposes. It is beyond the scope of this chapter to address, even in the most general way, all of the types of tests, rating scales, and the instrumentation that might be employed in the therapeutic environment. Instead, the focus here will be on general classes of instrumentation that have the greatest applicability in the service of the therapeutic endeavor. To a limited extent, specific examples of such instruments will be presented. This will be followed by a discussion of criteria and considerations that will assist the clinician in selecting the best instrumentation for his or her intended purposes.

4.18.4.1 Types of Instrumentation for Therapeutic Assessment

The instrumentation required for any therapeutic application will depend on: (i) the general purpose(s) for which the assessment is being conducted, and (ii) the level of informational detail that is required for those purpose(s). Generally, one may classify the types of instrumentation that would serve the purpose(s) of the therapeutic assessment into one of four general categories. As mentioned above, other types of instrumentation are frequently used in clinical settings for therapeutic purposes. However, the present discussion will be limited to those more commonly used by a wide variety of clinical psychologists in their day-to-day practices.

4.18.4.1.1 Psychological/psychiatric symptom measures

Probably the most frequently used instrumentation for several therapeutic purposes are measures of psychopathological symptomatology. Besides the fact that these are the types of instruments on which the majority of the clinician's psychological assessment training has probably been focused, they were developed to assess the problems that typically prompt people to seek treatment.

There are several subtypes of these measures of psychological/psychiatric symptomatology. The first is the comprehensive multidimensional measure. This is typically a lengthy, multiscale instrument that measures and provides a graphical profile of the patient on several types of psychopathological symptom domains (e.g., anxiety, depression) or disorders (schizophrenia, antisocial personality). Also, summary indices sometimes are available to provide a more global picture of the individual with regard to his or her psychological status or level

of distress. Probably the most widely used and/or recognized of these measures are the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1951) and its restandardized revision, the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), the Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, 1994), and the Personality Assessment Inventory (PAI; Morey, 1991).

Multiscale instruments of this type can serve a variety of purposes that facilitate therapeutic efforts. They may be used upon initial contact with the patient to screen for the need for service and, at the same time, yield information that is useful for treatment planning. Indeed, some such instruments (e.g., the MMPI-2) may make available supplementary, content-related, and/or special scales that are designed to assist the user in addressing specific treatment considerations (e.g., low motivation for treatment). Other multiscale instruments might be useful in identifying specific problems that may be unrelated to the patient's chief complaints (e.g., low self-esteem). They can also be administered at numerous times during the course of treatment to monitor the patient's progress toward achieving established goals and to assist in determining what adjustments (if any) must be made to the clinician's approach. In addition, use of the instrument in a pre- and post-treatment fashion provides information related to the outcomes of the treatment. Data obtained in this fashion can be analyzed with results from other patients to evaluate the effectiveness of an individual therapist as well as an organization.

Abbreviated multidimensional measures are quite similar to the comprehensive multidimensional measure in many respects. First, by definition, they contain multiple scales for measuring a variety of symptom domains and/or disorders. They also may allow for the derivation of an index of the patient's general level of psychopathology or distress. In addition, they may be used for screening, treatment planning and monitoring, and outcomes assessment purposes just like the comprehensive instruments. The distinguishing feature of the abbreviated instrument is its length. Again, by definition, these instruments are relatively short, and easy to administer and (usually) score. Their brevity does not allow for an in-depth assessment of the patient and his or her problems, but this is not what these instruments were designed to do.

Probably the most widely used of these brief instruments are Derogatis' family of symptom checklist instruments. These include the original Symptom Checklist-90 (SCL-90; Derogatis, Lipman, & Covi, 1973) and its revision, the

SCL-90-R (Derogatis, 1983). Both of these instruments contain a checklist of 90 psychological symptoms, most of which score on the instruments' nine symptom scales. For each of these instruments an even briefer version has been developed. The first is the Brief Symptom Inventory (BSI; Derogatis, 1992), which was derived from the SCL-90-R. In a health care environment that is cost-conscious and unwilling to make too many demands on patient time, this 53-item instrument is gaining popularity over its longer and more expensive 90-item parent instrument. Similarly, a brief form of the original SCL-90 has been developed. Titled the Symptom Assessment-45 Questionnaire (SA-45; Strategic Advantage, Inc., 1996), its development did not follow Derogatis' approach to the development of the BSI; instead, cluster analytic techniques were used to select five items each for assessing each of the nine symptom domains found on the three Derogatis checklists.

The major strength of the abbreviated multi-scale instruments is their ability to broadly and very quickly survey several psychological symptom domains and/or disorders relative to the patient. Its value is most clearly evident in settings where both the time and dollars available for assessment services are quite limited. These instruments provide a lot of information quickly. Because of their brevity, they are much more likely to be completed by patients than their lengthier comprehensive counterparts. This last point is particularly important if one is interested in monitoring treatment or assessing outcomes, both of which require at least two or more assessments to obtain the desired information.

4.18.4.1.2 Measures of general health status and role functioning

During the past decade, there has been an increasing interest in the assessment of health status in health care delivery systems. Initially, this interest was shown mostly by those organizations and settings focusing primarily on the treatment of physical diseases and disorders. Within recent years, behavioral health care providers have recognized the value in assessing the patient's general level of health. It is important to recognize that the term "health" means more than just the absence of disease or debility; it also implies a state of well-being throughout the individual's physical, psychological, and social spheres of existence (World Health Organization [WHO], 1948). Dickey and Wagenaar (1996) point out how this view of health recognizes the importance of eliciting the patient's point of view in assessing

health status. They also point to similar conclusions reached by Jahoda (1958) specific to the area of mental health. Here, an individual's self-assessment relative to how he or she feels they should be is an important component of "mental health."

Measures of health status and physical functioning can be classified into one of two groups: generic and condition-specific. Probably the most widely used and respected generic health status measures are the 36-item Medical Outcomes Study Short Form Health Scale (SF-36; Ware & Sherbourne, 1992; Ware, Snow, Kosinski, & Gandek, 1994) and the 39-item Health Status Questionnaire 2.0 (HSQ; Health Outcomes Institute, 1993; Radosevich, Wetzler, & Wilson, 1994). Aside from the minor variations in the scoring of one of the instruments' scales (i.e., Bodily Pain) and the HSQ's inclusion of three depression screening items, the two measures essentially are identical. Each assesses eight dimensions of health, four addressing mental health-related constructs and four addressing physical health-related constructs, that reflect the WHO concept of "health."

Role functioning has recently gained attention as an important variable to address in the course of assessing the impact of a physical or mental disorder on an individual. In devising a treatment plan and monitoring progress over time, it is important to know how the person's ability to work, perform daily tasks, or interact with others is affected by the disorder. The SF-36 and HSQ both address these issues with scales designed for this purpose.

Responding to concerns that even these relatively brief objective measures are too lengthy for regular administration in clinical and research settings, 12-item, abbreviated versions of each have been developed. The SF-12 (Ware, Kosinski, & Keller, 1995) was developed for use in large scale, population-based research where the monitoring of health status at a broad level is all that is required. Also, a 12-item version of the HSQ, the HSQ-12 (Radosevich & Pruitt, 1996), was developed for similar uses. Interestingly, given that the two abbreviated versions were derived from essentially the same instrument, there is only a 50% item overlap between the two shortened instruments. Both instruments are relatively new but the data supporting their use that has been gathered up to 1997 is promising.

Condition-specific health status and functioning measures have been utilized for a number of years. Most have been developed for use with physical rather than mental disorders, diseases, and conditions. However, condition-specific measures of mental health

status and functioning are beginning to appear. A major source of this type of instrument is the Minnesota-based Health Outcomes Institute (HOI), a successor to the health care think tank InterStudy. In addition to the HSQ and the HSQ-12, HOI serves as the distributor/clearing-house for the condition-specific “technology of patient experience (TyPE) specifications.” The available TyPEs that would be most useful to clinical psychologists and other behavioral health care practitioners include those developed by a team of researchers at the University of Arkansas Medical Center for use with depressive, phobic, and alcohol and substance disorders. TyPEs for other specific psychological disorders are currently under development at the University of Arkansas for distribution through HOI.

4.18.4.1.3 *Quality of life measures*

In their brief summary of this area, Andrews, Peters, and Teesson (1994) indicate that most of the definitions of “quality of life” (QOL) describe a multidimensional construct encompassing physical, affective, cognitive, social, and economic domains. Objective measures of QOL focus on environmental resources required to meet one’s needs and can be completed by someone other than the patient. The subjective measures of QOL assess the patient’s satisfaction with the various aspects of his or her life and thus must be completed by the patient.

Andrews et al. (1994) draw other distinctions in the QOL arena. One has to do with the differences between QOL and health-related quality of life, or HRQL, and (similar to the case with health status measures) the other has to do with the distinction between generic and condition-specific measures of QOL. QOL measures differ from HRQL measures in that the former assess the whole “fabric of life,” while the latter assess quality of life as it is affected by a disease or disorder, or by its treatment. Generic measures are designed to assess aspects of life that are generally relevant to most people; condition-specific measures are focused on aspects of the lives of particular disease/disorder populations. However, as Andrews et al. point out, generic and condition-specific QOL measures tend to overlap quite a bit.

4.18.4.1.4 *Service satisfaction measures*

With the exploding interest in assessing the outcomes of treatment for the patient, it is not surprising to see an accompanying interest in assessing the patient’s and, in some instances, the patient’s family’s satisfaction with the services received. In fact, many professionals

and organizations equate satisfaction with outcomes and frequently consider it the most important outcome. In a recent survey of 73 behavioral health care organizations, 71% of the respondents indicated that their outcomes studies included measures of patient satisfaction (Pallak, 1994).

Although some view service satisfaction as an outcome, it is this author’s contention that it should not be classified as such. Rather, it should be considered a measure of the overall therapeutic process, encompassing the patient’s (and at times, others’) view of how the service was delivered, the capabilities and attentiveness of the service provider, the benefits of the service (if any), and any of a number of other selected aspects of the service he or she received. Patient satisfaction surveys don’t answer the question “What was the result of the treatment rendered to the patient”; they do answer the question “How did the patient feel about the treatment he or she received?” Thus, they serve an important program evaluation/improvement function.

The number of questionnaires that are currently being used to measure patient satisfaction is countless. This reflects the attempts of individual health care organizations to develop customized measures that assess variables important to their particular needs, which in turn reflects a response to outside demands to “do something” to demonstrate the effectiveness of their services. Often, this “something” has not been evaluated to determine its basic psychometric properties. As a result, there exists numerous options that one may choose from, but very few that actually have demonstrated their validity and reliability as measures of service satisfaction.

Fortunately, there are a few instruments that have been investigated for their psychometric integrity. Probably the most widely used and researched patient satisfaction instrument designed for use in behavioral health care settings is the eight-item version of the Client Satisfaction Questionnaire (CSQ-8; Attkisson & Zwick, 1982; Nguyen, Attkisson, & Stenger, 1983). The CSQ-8 was derived from the original 31-item CSQ (Larsen, Attkisson, Hargreaves, & Nguyen, 1979), which also yielded two longer 18-item alternate forms, the CSQ-18A and CSQ-18B (LeVois, Nguyen, & Attkisson, 1981). The more recent work of Attkisson and his colleagues at the University of California at San Francisco is the Service Satisfaction Scale-30 (SSS-30; Greenfield & Attkisson, 1989), a 30-item multifactorial scale that yields information regarding different aspects of satisfaction with mental health service, such as perceived outcome and manner and skill of the clinician.

4.18.4.2 Guidelines for Instrument Selection

Regardless of the type of instrument one might consider using in the therapeutic environment, many clinical psychologists frequently must choose between many product offerings. But what are the general criteria for the selection of any instrument for psychological assessment? What should guide the clinician's selection of an instrument for a specific therapeutic purpose? As part of their training, clinical psychologists and professionals from related psychological specialties have been educated about the important psychometric properties that should be considered when determining the appropriateness of an instrument for its intended use. However, this is just one of several issues that should be taken into account in an evaluation of a specific instrument for a specific therapeutic use. The guidance that has been offered by experts with regard to instrument selection is worth noting here.

4.18.4.2.1 *National Institute of Mental Health criteria*

Probably the most thorough and clinically relevant guidelines for the selection of psychological assessment instruments comes from the National Institute of Mental Health (NIMH) supported work of Ciarlo, Brown, Edwards, Kiresuk, and Newman (1986). A synopsis of Newman and Ciarlo's (1994) updated summary of this NIMH work is presented here. Note that the criteria discussed below were originally developed for use in evaluating instruments for outcomes assessment purposes. However, most have relevance to the selection of instrumentation used for the other therapeutic assessment purposes described above. Exceptions and qualifications with regard to this issue will be noted when appropriate.

Newman and Ciarlo (1994) describe 11 criteria for the selection of outcomes assessment instruments, each of which can be grouped into one of five types of consideration. The first consideration is that of applicability. The issue here is the relevance of the instrument to the target population. The instrument should assess those problems, symptoms, characteristics, and so on, that are common to the group to whom the instrument will be administered. The more heterogeneous the population, the more chance that modifications will be required and that these will alter the standardization and psychometric integrity of the instrument. Another applicability issue to consider when the instrument is to be used for outcomes assessment purposes is its indepen-

dence from the type of treatment to be offered to the population.

The second set of general considerations is that of methods and procedures (Newman & Ciarlo, 1994). Several selection criteria are related to this group. The first is that administration of the instrument is simple and easily taught. Generally, this is more of an issue with clinician-rating scales than self-report scales. In the case of rating scales, concrete examples, or objective referents, at each rating level should be provided to the user. Next, the instrument should allow input not only from the patient but also from other sources (e.g., the clinician, collaterals). The benefits of this include the opportunities to obtain a feel for the patient from many perspectives, to validate reported findings and observations, and to promote honesty in responding from all sources (given that all parties will know that others will also be providing input). The final methods and procedures criterion, though not necessarily as important for the instrument being used for screening or treatment planning purposes, is that the instrument provide information relevant to understanding how the treatment may have effected change in the individual.

Newman and Ciarlo's (1994) third set of considerations have to do with the psychometric strengths of the instruments. According to the NIMH panel of experts, outcomes measures should: (i) meet the minimum psychometric standards for reliability (including internal consistency, test-retest reliability, and as appropriate, interrater reliability) and validity (content, construct, and concurrent validity); (ii) be difficult to "fake bad" or "fake good"; and (iii) be free from response bias and not reactive or sensitive to factors unrelated to the constructs that are being measured (e.g., physical settings, behavior of the treatment staff). These criteria obviously also apply to other psychological instruments used for purposes other than outcomes assessment. However, for outcomes assessment purposes, the instrument also must be sensitive to change related to treatment.

The fourth group of considerations concerns the cost of the instruments. Newman and Ciarlo (1994) point out that the answer to the question of how much one should spend on assessment instrumentation and associated costs (e.g., staff time for administering, scoring, processing, and analyzing the data) will depend on how important the data gathered is to assuring a positive return on the functions they support. In the context of the NIMH undertaking, Newman and Ciarlo felt that the data obtained through treatment outcomes assessment would support screening/treatment planning, efforts in quality assurance and program

evaluation, cost containment/utilization review activities, and revenue generation efforts. However, that may be considered the ideal. At this point, the number and nature of the purposes that would be supported by the obtained data will depend on the individual organization. The more purposes the data can serve, the less costly the instrumentation is likely to be, at least from a value standpoint. In terms of actual costs, Ciarlo et al. (1986) estimated that 0.5% of an organization's total budget would be an affordable amount for materials, staff training, data collection, and processing costs related to outcomes assessment. However, one should be mindful that the recommendation was made in 1986 and may not reflect changes in policies, requirements, and attitudes related to the use of psychological assessment instruments since that time.

The final set of considerations in instrument selection has to do with the utility of the instrument. Four criteria related to utility are posited by Newman and Ciarlo (1994). First, the scoring procedures and the manner in which the results are presented should be comprehensible to all with a stake in the treatment of the organization's patients. This would not only include the patient, his or her family, the organization's administrative staff and other treatment staff, but also third-party payers and (in the case of outcomes assessment or program evaluation) legislative and administrative policy makers. Related to this is the criterion that the results of the instrument be easily interpreted by those with a stake in them. Another utility-related criterion is that the instrument should be compatible with a number of clinical practices and theories that are employed in the behavioral health care arena. This should allow for a greater range of test applicability and greater acceptance by the various stakeholders in the patient's treatment.

Another important aspect of utility is that "the instrument support[s] the clinical processes of a service with minimal interference" (Newman & Ciarlo, 1994, p. 107). There are two issues here. The first has to do with whether the instrument can support the screening, planning, and/or monitoring activities in addition to the outcomes assessment activities. In other words, are multiple purposes served by the instrument's results? The second issue is one that has to do with the extent to which the organization's staff is burdened with the collection and processing of assessment data. How much will the assessment process interfere with the daily work flow of the organization's staff? Equally important is whether the benefits that accrue justify the cost of implementing an assessment program for whatever purpose(s).

4.18.4.2.2 Other criteria and considerations

Although the work of Ciarlo and his colleagues provides more extensive instrument selection guidelines than most, others who have addressed the issue have arrived at recommendations that serve to reinforce and/or complement those found in the NIMH document. For example, Gavin Andrews' work in Australia has led to significant contributions to the body of outcomes assessment knowledge. As part of this, Andrews et al. (1994) have identified six general "qualities of consumer outcome measures" that are generally in concordance with those from the NIMH study. First, the measure should meet the criterion of applicability. In other words,

it should address dimensions which are important to the consumer (symptoms, disability, and consumer satisfaction) and useful for the clinician in formulating and conducting treatment, yet the measure should be one which can have its data aggregated in a meaningful way so that the requirements of management can be addressed. (p. 30)

Multidimensional instruments yielding a profile of scores on all dimensions of interest are viewed as a means of best serving the interests of all concerned.

Acceptability, that is, being both brief and user-friendly, is another desirable quality identified by Andrews et al (1994). Closely associated with this is the criterion of practicality. It might be viewed as a composite of those NIMH criteria related to matters of cost, ease of scoring and interpretation, and training in the use and interpretation of the measure. Again in agreement with the NIMH work, the final three criteria identified by Andrews et al. relate to reliability, validity, and sensitivity to change. With regard to reliability, Andrews et al. specify what they consider to be the minimum levels of acceptable internal consistency reliability (0.90 for long tests), interrater reliability (0.40), and construct and criterion validity (0.50). They also stress the importance of an instrument's face validity in helping to ensure cooperation from the patient, and of self-report instruments having multiple response options (rather than just "yes/no" options) for increasing sensitivity of an instrument to small but relevant changes in the patient's status over time.

In Ficken's (1995) discussion of the role of assessment in an MCO environment, he concludes that the difficulties clinicians are experiencing in demonstrating the utility of psychological assessment to payers lies in the fact the instruments and objectives of traditional psychological assessment are not in synch

with the needs of MCOs. The solution to the problem appears simple:

the underlying objectives of testing must be aligned with the values and processes of MCOs. In short, this means identifying decision points in managed care processes that could be improved with objective, standardized data. There are two avenues in which these can be pursued: through facilitation/objectification of clinical-decision processes and through outcome assessment. (p. 12)

In general, Ficken (1995) sees opportunities in areas that this author has previously identified as screening, treatment planning, and outcomes assessment, specifically in the areas of primary medical care and behavioral health care (see below). Requirements of instruments used for screening were noted to include:

(i) high levels of sensitivity and specificity to diagnostic criteria from the *Diagnostic and statistical manual of mental disorders* (4th ed., *DSM-IV*; American Psychiatric Association, 1994) or the most up-to-date version of the *International classification of diseases* (*ICD*);

(ii) a focus on hard-to-detect (in a single office visit) but treatable disorders that are associated with imminent harm to self or others, significant suffering, and a decrease in productivity;

(iii) an administration time of no more than 10 minutes; and

(iv) an administration protocol that easily integrates into the organization's work flow. Cases testing "positive" on the screener would be administered one or more "second-tier" instrument(s) to establish severity and a specific diagnosis. Ficken feels that if they are to be accepted by MCOs, these second-tier instruments should meet the requirements of screeners and either specify or rule out a diagnosis.

According to Ficken (1995), successful outcomes assessment instruments also must possess certain qualities. Because the areas most important to assess for outcomes measurement purposes are symptom reduction, level of functioning, quality of life and patient satisfaction, the instrument should (i) focus on one of these areas, (ii) be brief, (iii) meet "traditional standards" for validity and reliability, and (iv) be sensitive to clinical change.

Based on the work of Vermillion and Pfeiffer (1993), Burlingame, Lambert, Reisinger, Neff, and Mosier (1995) recommended four criteria for the selection of outcomes measures. The first is acceptable "technical features," that is, validity and reliability. Specifically, these authors recommended that instruments have an internal consistency of at least 0.80, test-retest reliability of at least 0.70, and concurrent

validity of at least 0.50. The second criterion is "practicality features." These include brevity, ease of administration and scoring, and simplicity of the reporting of results. Third, the instrumentation should be "suitable" for the patients that are seen within the setting. Thus, because of the nature of most presenting problems in mental health settings, it should assess symptomatology and psychosocial functioning. The fourth criterion is sensitivity to "meaningful" change over time, allowing for a differentiation of *symptomatic* change from interpersonal/social role *functional* change.

Schlosser (1995) proposed a rather nontraditional view of "outcomes assessment." In what he refers to as a "patient-centric" view, assessment information is gathered and used during the course of therapy to bring about change during therapy, not after therapy has ended. Essentially, this equates to what this author has referred to above (and discusses in more detail below) as treatment monitoring. In this model, Schlosser feels that this type of assessment requires "elements regarding very specific, theoretically derived, empirically validated areas of functioning" (Schlosser, 1995, p. 66). These would involve the use of both illness and well-being measures that assess the patient on emotional, mental/cognitive, physical, social, life direction, and life satisfaction dimensions.

Many of Schlosser's (1995) considerations for selection of such measures are not unique (i.e., having "acceptable" levels of reliability and validity, brief, low-cost, and sensitive). However, for the purposes described Schlosser also indicates that they should also: (i) have "paradigmatic sensibility" (i.e., key words have the same meaning across instruments); (ii) be designed for repeated administration for feedback or self-monitoring purposes; and (iii) provide actionable information.

In addition to some already mentioned criteria (acceptable validity, reliability, affordability, ease of administration, and ease of data entry and analysis), Sederer et al. (1996) discuss other considerations that warrant attention in selecting outcomes measures for specific situations. These include automation capabilities related to availability of software for data analysis and reporting, compatibility with the organization's existing information system, and the ability to enter data via an optical scanner. They also provide advice that should help guide the user in selecting the appropriate instrumentation:

A plan should be developed that addresses the following questions: which patients will be included in the study? What outcomes will be most

effected by the treatment? When will the outcomes be measured? Who is going to read (and use) the information provided by the outcomes study? The more specific the answer to these questions, the better the choice of outcome instrument. (p. 4)

This and other recommendations would appear equally applicable when selecting instruments for other therapeutic assessment purposes.

One final set of criteria should be considered in the light of the following section on screening. Screening for the likelihood of the presence of disorders or for the need for additional assessment requires considerations that do not necessarily apply to instruments when they are used for the other therapeutic assessment purposes addressed in this chapter. A major one here is a specific consideration relative to a screener's criterion validity. Although broadly encompassed by the construct of "validity" that was previously discussed, it demands particular attention when evaluating instruments for screening purposes. What is being referred to here is the instrument's classification accuracy or efficiency.

Classification efficiency is usually expressed in terms of the following statistics: *sensitivity*, that is, the proportion of those individuals with the characteristic of interest who are accurately identified as such; *specificity*, that is, the proportion of individuals not having the characteristic of interest who are accurately identified as such); *positive predictive power*, which is the proportion of a population identified by the instrument as having the characteristic who actually do have the characteristic, and *negative predictive power*, which is the proportion of a population identified by the instrument as not having the characteristic who actually do not have the characteristic. This information can provide the clinical psychologist and other evaluators with empirically based information that is useful in the type of decision-making requiring the selection of one of two choices. The questions answered are typically those of the "yes/no" type, such as "Is the patient depressed or not?" or "Does the patient have a psychological problem significant enough to require treatment?" The reader is referred to Baldessarini, Finkelstein, and Arana (1983) for a discussion of issues related to the use of these statistics.

In evaluating these statistics, one must consider a few very important issues. One is the degree to which the clinician is willing to accept false-positives or false-negatives. This will be a function of the importance of maximizing the correct identification of those with the particular characteristic of interest vs. the importance of maximizing the correct

identification of those not having the characteristic vs. the importance of optimizing the identification of both groups. This in turn will be dependent on the cutoff score recommended by the developer of the instrument and/or the efficiency values that are available when other cutoff scores are applied. These and related issues are discussed more extensively in the next section.

4.18.5 PSYCHOLOGICAL ASSESSMENT AS A TOOL FOR SCREENING

One of the most significant ways in which psychological assessment can contribute to the development of an economic and efficient behavioral health care delivery system is by using it to screen potential patients for need for behavioral health care services, and/or to determine the likelihood that the problem being screened is a particular disorder of interest. Probably the most concise, informative treatment of the topic of the use of psychological tests in screening for behavioral health care disorders is provided by Derogatis and DellaPietra (1994). In this work, these authors turn to the Commission on Chronic Illness (1987) to provide a good working definition of health care screening in general, that being:

the presumptive identification of unrecognized disease or defect by the application of tests, examinations or other procedures which can be applied rapidly to sort out apparently well persons who probably have a disease from those who probably do not. (Commission on Chronic Illness, 1987, p. 45)

Derogatis and DellaPietra (1994) further clarify the nature and the use of screening procedures, stating that:

the screening process represents a relatively unrefined sieve that is designed to segregate the cohort under assessment into "positives" who presumptively have the condition, and "negatives" who are ostensibly free of the disorder. Screening is not a diagnostic procedure *per se*. Rather, it represents a preliminary filtering operation that identifies those individuals with the highest probability of having the disorder in question for subsequent specific diagnostic evaluation. Individuals found negative by the screening process are not evaluated further. (p. 23)

The most important aspect of any screening procedure is the efficiency with which it can provide information useful to clinical decision-making. In the area of clinical psychology, the most efficient and thoroughly investigated screening procedures involve the use of psycho-

logical assessment instruments. As implied by the foregoing, the power or utility of a psychological screener lies in its ability to determine, with a high level of probability, whether the respondent does or does not have a particular disorder or condition, or whether he or she is or is not a member of a group with clearly defined characteristics. In daily clinical practice, the most commonly used screeners are those designed specifically to identify some aspect of psychological functioning or disturbance or provide a broad overview of the respondent's point-in-time mental status. Examples of problem-specific screeners include the Beck Depression Inventory (BDI; Beck, Rush, Shaw, & Emery, 1979) and the State-Trait Anxiety Inventory (STAI; Spielberger, 1983). Examples of screeners for more generalized psychopathology or distress include the SA-45 and BSI.

4.18.5.1 Research-based Use of Psychological Screeners

The establishment of a system for screening for a particular disorder or condition involves determining what it is one wants to screen in or screen out, at what level of probability one feels comfortable in making that decision, and how many incorrect classifications or what percentage of errors one is willing to tolerate. Once it is decided what one wishes to screen for, one then must turn to the instrument's classification efficiency statistics, that is, sensitivity, specificity, positive predictive power (PPP), and negative predictive power (NPP), for the information necessary to determine whether a given instrument is suitable for the intended purpose(s).

Recall that sensitivity refers to the proportion of those with the characteristic of interest who are accurately identified as such by an instrument or procedure, while specificity refers to the proportion of those not having the characteristic of interest who are accurately identified. The cutoff score, index value, or other criterion used for classification can be adjusted to maximize either sensitivity or specificity. However, maximization of one will necessarily result in a decrease in the other, thus increasing the percentage of false-positives (with maximized sensitivity) or false-negatives (with maximized specificity). Stated differently, false-positives will increase as specificity decreases, while false-negatives will increase as sensitivity decreases (Elwood, 1993).

Another approach is to optimize both sensitivity and specificity, thus yielding a fairly even balance of true positives and true negatives. Although optimization might seem to be the

preferable approach in all instances, there are situations in which a maximization approach is more desirable. For example, a psychiatric hospital with an inordinately high rate of inpatient suicide attempts begins to employ a screener designed to help identify patients with suicide potential as part of its admission procedures. The hospital adjusts the classification cutoff score to a level that identifies all suicidal patients in the screener's normative group. This cutoff score is then applied to all patients being admitted to the hospital for the purpose of identifying those requiring an extensive evaluation for suicide potential. This not only increases the number of true positives, but it also decreases the specificity and increases the number of false positives. However, the trade-off of identifying more suicidal patients early on with having more nonsuicidal patients receiving suicide evaluations would appear worthwhile for the hospital's purposes. Similarly, in other instances, maximization of specificity may be the preferred approach. For example, an MCO might wish to use a measure of overall level of psychological distress to identify those covered lives that are not in need of behavioral health care services. Sensitivity will decrease but, for the MCO's purposes, this might be quite acceptable.

Hsiao, Bartko, and Potter (1989) note that "a diagnostic test will not have a unique sensitivity and specificity. Instead, for each diagnostic test, the relationship between sensitivity and specificity depends on the cutoff point chosen for the test" (p. 665). The effect of employing individual classification cutoff points can be presented via the use of receiver operating characteristic (ROC) curves. These curves are nothing more than a plotting of the resulting true positive rate (sensitivity) against the false positive rate for each cutoff score that might be employed with a test used for classification purposes. The plotting allows for a graphical representation of what may be gained and/or lost by shifting cutoff scores. The resulting area underneath the curve provides an indication of how well the test performs. Development of ROC curves from available data for a test being considered for screening purposes is recommended. The reader is referred to Hsiao et al. and Metz (1978) for a more detailed discussion of ROC curves and their use.

In day-to-day clinical work, an instrument's PPP and NPP can provide information that is more useful than sensitivity and specificity. As Elwood (1993) has pointed out,

Although sensitivity and specificity do provide important information about the overall performance of a test, their limitation in classifying

individual subjects becomes evident when they are considered in terms of conditional probabilities. Sensitivity is $P(+/d)$, the probability (P) of a positive test result (+) given that the subject has the target disorder (d). However, the task of the clinicians in assessing individual patients is just the opposite: determining $P(d/+)$, the probability that a patient has the disorder given that he or she obtained an abnormal test score. In the same way, specificity expresses $P(-/-d)$, the probability that a patient will have a negative test result given that he or she does not have the disorder. Here again, the task confronting the clinician is usually just the opposite: determining $P(-d/-)$, the probability that the patient does not have the disorder given a negative test result. (p. 410)

A note of caution is warranted when evaluating the two predictive powers of a test. Unlike sensitivity and specificity, both PPP and NPP are affected and change according to the prevalence or base rate at which the condition or characteristic of interest (i.e., that which is being screened by the test) occurs within a given setting. As Elwood (1993) reports, the lowering of base rates results in lower PPPs while increasing base rates result in higher PPPs. The opposite trend is true for NPPs. He notes that this is an important consideration because clinical tests are frequently validated using samples in which the prevalence rate is 0.50, or 50%. Thus, it is not surprising to see a test's PPP drop in "real-life" applications where the prevalence is lower.

Derogatis and DellaPietra (1994) indicate that a procedure referred to as "sequential screening" may provide at least a partial solution to the limitations or other problems that low base rates may pose for the predictive powers of an instrument. Sequential screening essentially involves the administration of two screeners, each of which measures the condition of interest, and two-phase screening. In the first phase, one screener is administered to the low base rate population. The purpose of this is to identify those individuals without the condition, thus requiring relatively good specificity. These individuals are eliminated from involvement in the second phase, resulting in an increase in the prevalence of the condition among those who remain. This group is then administered another screener of equal or better sensitivity. With the increased prevalence of the condition in the remaining group, the false positive rate will be much lower. As Derogatis and DellaPietra point out,

Sequential screening essentially zeros in on a high-risk subgroup of the population of interest by virtue of a series of consecutive sieves. These have

the effect of eliminating from consideration individuals with low likelihood of having the disorder, and simultaneously raising the base rate of the condition in the remaining sample. (p. 45)

In summary, PPP and NPP can provide information that is quite valuable to those making important clinical decisions, such as determining need for behavioral health care services, assigning diagnoses, or determining appropriate level of care. However, these users must be cognizant of the manner in which the predictive powers may change with the population to which the test or procedure is applied.

4.18.5.2 Implementation of Screeners into the Daily Work Flow of Service Delivery

The utility of a screening instrument is only as good as the degree to which it can be integrated into an organization's daily regimen of service delivery. This, in turn, depends on a number of factors. The first is the degree to which the administration and scoring of the screener is quick and easy, and the amount of time required to train the provider's staff to successfully incorporate the screener into their day-to-day activities.

The second factor relates to its use. Here, the screener is not used for anything other than determining the likelihood that the patient does or does not have the specific condition or characteristic the instrument is designed to assess. Use for any other purpose (e.g., assigning a diagnosis based solely on screener results, determining the likelihood of the presence of other characteristics) only serves to undermine the integrity of the instrument in the eyes of staff, payers, and other parties with a vested interest in the screening process.

The third factor has to do with the ability of the provider to act on the information. It must be clear how the clinician should proceed based on the information available.

The final factor is staff acceptance and commitment to the screening process. This comes only with a clear understanding of the importance of the screening, the usefulness of the obtained information, and how the screening process is to be incorporated into the organization's business flow.

Ficken (1995) provides an example of how screeners can be integrated into an assessment system designed to assist primary care physicians to identify patients with psychiatric disorders. This system (which also allows for the incorporation of practice guidelines) seems to take into account the first three utility-related factors listed above. It begins with the administration of a screener that is highly

sensitive and specific to *DSM*- or *ICD*-related disorders. Ficken indicates that screeners should require no more than 10 minutes to complete, and that “their administration must be integrated seamlessly into the standard clinical routine” (p. 13). Somewhat similarly to the sequence described by Derogatis and DellaPietra (1994), positive findings would lead to a second level of testing. Here, another screener that meets the same requirements as those for the first screener and also affirms or rules out a diagnosis would be administered. Positive findings would lead to additional assessment for treatment planning purposes. Consistent with standard practice, Ficken recommends confirmation of screener findings by a qualified psychologist or physician.

4.18.6 PSYCHOLOGICAL ASSESSMENT AS A TOOL FOR TREATMENT PLANNING

The administration of screeners is only one way in which psychological assessment can serve as a valuable tool for treatment planning. However, many would argue that it is the most limited way in which this tool can be used for planning a course of treatment. When employed by a trained clinician, psychological assessment can provide information that can greatly facilitate and enhance the planning of the therapeutic intervention for the individual patient.

The importance of treatment planning has received significant attention during recent years. The reasons for this were summarized previously by this author (Maruish, 1990) as follows:

Among important and interrelated reasons ... [are] concerted efforts to make psychotherapy more efficient and cost effective, the growing influence of “third parties” (insurance companies and the federal government) that are called upon to foot the bill for psychological as well as medical treatments, and society’s disenchantment with open-ended forms of psychotherapy without clearly defined goals. (p. iii)

The role that psychological assessment can play in planning a course of treatment for behavioral health care problems is significant. Butcher (1990) indicated that information available from instruments such as the *MMPI-2* can not only assist in identifying problems (see above) and establishing communication with the patient (see below), it can also help ensure that the plan for treatment is consistent with the patient’s personality and external resources. In addition, psychological assessment may reveal

potential obstacles to therapy, areas of potential growth, and problems of which the patient may not be consciously aware. Moreover, both Butcher and Appelbaum (1990) viewed testing as a means of quickly obtaining a second opinion. Other benefits of the results of psychological assessment, identified by Appelbaum, include assistance in identifying patient strengths and weaknesses, identification of the complexity of the patient’s personality, and establishment of a reference point or guide to refer to during the therapeutic episode.

The types of information that can be derived from patient assessment and the manner in which it is applied for this purpose are quite varied—a fact that will become evident below. Nevertheless, Strupp (see Butcher, 1990) probably provided the best summary of the potential contribution of psychological assessment to treatment planning, stating that “careful assessment of patient’s personality resources and liabilities is of inestimable importance. It will predictably save money and avoid misplaced therapeutic effort; it can also enhance the likelihood of favorable treatment outcomes for suitable patients” (pp. v–vi).

4.18.6.1 Assumptions About Treatment Planning

The introduction to this section presented a broad overview of ways in which psychological assessment can assist in devising and successfully implementing plans of treatment for behavioral health care patients. These and other benefits will be discussed in greater detail below. However, it is important to first clarify what treatment planning is and some of the general, implicit assumptions that one typically can make about this important therapeutic activity.

For the purpose of this discussion, the term “treatment planning” indicates that part of a therapeutic episode in which the treatment provider develops a set of goals for an individual presenting with behavioral health care problems, and outlines the specific means by which he/she or other resources will assist the patient in achieving those goals in the most efficient manner. General assumptions underlying the treatment planning process are as follows.

(i) The patient is experiencing behavioral health problems that have been identified either by themselves or by another party. Common external sources of problem identification include the patient’s spouse, parent, teacher, employer, and the legal system.

(ii) The patient experiences some degree of internal and/or external motivation to eliminate

or reduce the identified problems. An example of external motivation to change is the potential loss of job or marriage if problems are not resolved.

(iii) The goals of treatment are tied either directly or indirectly to the identified problems.

(iv) The goals of treatment have definable criteria for achievement, are indeed achievable by the patient, and are developed in collaboration with the patient.

(v) The prioritization of goals is reflected in the treatment plan.

(vi) The patient's progress toward achievement of the treatment goals can be tracked and compared against an expected path of improvement in either a formal or informal manner. This expected path of improvement may be based on the clinician's experience or (ideally) on objective data gathered on patients similar to the patient.

(vii) Deviations from the expected path of improvement will lead to a modification in the treatment plan, followed by subsequent monitoring to determine the effectiveness of the alteration.

These assumptions should not be considered exhaustive, nor are they reflective of what actually occurs in all situations. For example, some patients seen for therapeutic services may have no motivation to change. As may be seen in juvenile detention settings or in cases where children are brought to treatment by the parents, their participation in treatment is forced, and they may engage in intentional efforts to sabotage any therapeutic intervention. Also, it is likely that there are still clinicians who identify and prioritize treatment goals without the direct input of the patient. Nevertheless, the assumptions above represent this author's view of the aspects of treatment planning that have a direct bearing on the manner in which psychological assessment can best serve treatment planning efforts.

4.18.6.2 The Benefits of Psychological Assessment for Treatment Planning

As has already been touched upon, there are several ways in which psychological assessment can assist in the planning of treatment for behavioral health care patients. Following is a discussion of the more common and evident contributions that assessment can make to treatment planning efforts. These can be organized into four general categories: problem identification, problem clarification, identification of important patient characteristics, and monitoring of treatment progress.

4.18.6.2.1 Problem identification

Probably the most common use of psychological assessment in the service of treatment planning is for the purpose of problem identification. Often, the use of psychological testing *per se* is not needed to identify what problems the patient is experiencing. He or she will either tell the clinician directly without questioning, or they will readily admit to their problem(s) during the course of a clinical interview. However, this is not always the case.

The value of psychological testing becomes apparent in those cases where the patient is hesitant or unable to identify the nature of his or her problems. However, with a motivated and engaged patient who responds to items on a well validated and reliable test in an open and honest manner, the process of identifying what brought the patient to treatment also may be greatly facilitated. Cooperation shown during testing may be attributable to the nonthreatening nature of responding to questions presented on paper or a computer monitor (as opposed to those posed by another human being); the subtle, indirect, or otherwise nonthreatening nature of the questions (compared to those asked by the clinician); instrumentation that "casts a wider net" than the clinician in his or her interview with the patient; or any combination of these reasons.

In addition, the nature of some of the more commonly used psychological test instruments allows for the identification of secondary problems of significant severity that might otherwise be overlooked. Multidimensional inventories such as the MMPI-2 and the PAI are good examples of these types of instruments. Moreover, these instruments may be sensitive to other problems or patient traits or characteristics that may not necessarily be problems but which may exacerbate or otherwise contribute to the maintenance of the patient's problems.

Note that the type of problem identification described here is different from that conducted during screening (see above). Whereas screening is focused on determining the presence or absence of a single problem, problem identification generally takes a broader view and investigates the possibility of the presence of multiple problem areas. At the same time, there is also an attempt to determine the extent to which the problem area(s) affect the patient's ability to function.

4.18.6.2.2 Problem clarification

Psychological testing can often assist in the clarification of a known problem. Through tests designed for use with individuals presenting

problems similar to the patient's, aspects of identified problems can be elucidated. This will improve the patient's and clinician's understanding of the problem and likely lead to a better treatment plan. The three most important types of information that can be gleaned for this purpose are the severity of the problems, the complexity of the problems, and the degree to which the problems impair the patient's ability to function in one or more life roles.

The manner in which a patient is treated depends a great deal on the severity of his or her problem. In particular, severity has a great bearing on the setting in which the behavioral health care intervention is provided. Those whose problems are so severe that they are considered a danger to themselves or others more often than not are best suited for inpatient treatment, at least until dangerousness is no longer an issue. Similarly, problem severity may be a primary criterion for an evaluation for a medication adjunct to treatment. Severity also may have a bearing on the type of psychotherapeutic approach that is taken by the clinician. For example, it may be more productive for the clinician to take a supportive role with severe cases; all things being equal, a more confrontational approach may be more appropriate with patients with problems in the mild to moderate range of severity.

As alluded to above, the problems of patients seeking behavioral health care services are frequently multidimensional. Patient and environmental factors that play into the formation and maintenance of a problem, along with the latter's relationship with other problems, all contribute to its complexity. Knowing the complexity of the target problems is invaluable in devising an effective treatment plan. Again, multidimensional instruments or batteries of tests measuring specific aspects of psychological dysfunction serve this purpose well.

As with problem severity, knowledge of the complexity of a patient's psychological problems can help the clinician and patient in many aspects of treatment planning, including determination of appropriate setting, therapeutic approach, need for medication, and other matters on which important decisions must be made. However, possibly of equal importance and concern to the patient and outside parties (spouse, employer, school, etc.) is the extent to which these problems affect the patient's ability to function in his or her role as parent, child, employee, student, friend, and so on. Data gathered from the administration of measures of role functioning can provide information that not only clarifies the impact of the patient's problems and serves to establish role-specific goals, but also identifies other parties that may

serve as potential allies in the therapeutic process. In general, the most important role-functioning domains for assessment would be those related to work or school performance, interpersonal relationships, and activities of daily living (ADLs).

4.18.6.2.3 Identification of important patient characteristics

The identification and clarification of the patient's problems is of key importance in planning a course of treatment for the patient. However, there are numerous types of nonproblem-oriented patient information that can be useful in planning treatment and can be rather easily identified through the use of psychological assessment instruments. The vast majority of treatment plans are developed or modified with consideration of at least some of these other patient characteristics. The exceptions mostly are found with clinicians or programs that take a "one size fits all" approach to the treatment of general or specific types of disorders. It is beyond the scope of this chapter to provide an exhaustive list of what other types of information may be available to the clinician. However, a few are particularly worth mentioning.

Probably the most useful type of nonproblem-oriented information that can be gleaned from psychological assessment results is the identification of the patient characteristics or conditions that can serve as assets or areas of strength for the patient in working toward achieving the therapeutic goals. For example, Morey and Henry (1994) point to the utility of the PAI's Nonsupport scale in identifying whether the patient perceives an adequate social support network, this being a predictor of positive therapeutic progress. Other examples include "normal" personality characteristic information, such as that which can be obtained from Gough, McClosky, and Meehl's Dominance and Social Responsibility scales (1951, 1952) developed for use with the MMPI/MMPI-2. Greene (1991) indicates that those with high scores on the Dominance scale are described as "being able to take charge of responsibility for their lives. They are poised, self-assured, and confident of their own abilities" (p. 209). Gough and his colleagues interpreted high scores on the Social Responsibility scale as being indicative of individuals who, among other things, trust the world, are self-assured and poised, and stress the need for one to carry his or her share of duties. Thus, scores on these scales may reveal some important aspects of patient functioning that can be used in the service of affecting therapeutic change.

Similarly, knowledge of the patient's weaknesses or deficits may impact the type of treatment plan that is devised. Greene and Clopton (1994) provided numerous types of deficit-relevant information from the MMPI-2 Content Scales that have implications for treatment planning. For example, a clinically significant score ($T > 64$) on the Anger scale should lead one to consider the inclusion of training in assertiveness and/or anger control as part of the patient's treatment. On the other hand, uneasiness in social situations, as suggested by a significantly elevated score on either the Low Self-Esteem or Social Discomfort scale, suggests that a supportive approach to the intervention would be beneficial, at least initially.

Moreover, use of specially designed scales and procedures can provide information related to the patient's ability to become engaged in the therapeutic process. For example, the MMPI-2 Negative Treatment Indicators content scale developed by Butcher and his colleagues (Butcher, Graham, Williams, & Ben-Porath, 1989) may be useful in determining whether the patient is likely to be resistant to any form of "talk" therapy. Morey and Henry (1994) have supplied algorithms utilizing T scores for various PAI scales to make statements about the presence of positive characteristics, such as the presence of sufficient distress to motivate engagement in treatment, the ability to form a therapeutic alliance, and the capacity to utilize psychotherapy. The Therapeutic Reactance Scale (Dowd, Milne, & Wise, 1991) is yet another example of an instrument from which the clinician can be forewarned of potential resistance to therapeutic intervention.

Other types of patient characteristics that can be identified through psychological assessment have implications for the choice of the therapeutic approach and thus can contribute significantly to the treatment planning process. Beutler and his colleagues (Beutler & Clarkin, 1990; Beutler, Wakefield, & Williams, 1994; Beutler & Williams, 1995) have identified four patient characteristics that are thought to be important to matching patients and treatment approach for maximized therapeutic effectiveness. These include symptom severity, symptom complexity, coping style, and potential resistance to treatment. At different points in time, other patient variables also have been identified by these investigators as important considerations in the selection of the best treatment for a given patient. These include the problem-solving phase the patient has reached (Beutler & Clarkin, 1990), and subjective distress and social support (L.E. Beutler, personal communication, January 15, 1996).

Moreland (1996) points out how psychological assessment can assist in determining whether the patient deals with problems through internalizing or externalizing behaviors. All things being equal, internalizers would probably profit most from an insight-oriented approach rather than a behaviorally oriented approach. The reverse would be true for externalizers. In addition, cognitive factors also are important. Knowing that intelligence test results indicate an average or above IQ can assist the clinician in determining whether a patient will be able to benefit from a cognitive approach.

4.18.6.2.4 Monitoring of progress along the path of expected improvement

Information from repeated testing during the treatment process can help the clinician to determine if the treatment plan is appropriate for the patient at that particular point in time. Thus, many clinicians use psychological assessment to determine whether their patients are showing the expected improvement as treatment progresses. If not, adjustments can be made. These adjustments may reflect the need for a more intensive or aggressive treatment approach (e.g., increased number of psychotherapeutic sessions each week, addition of a medication adjunct) or for a less intensive approach (e.g., reduce or terminate medication, transfer from inpatient to outpatient care). Either way, this may require further retesting in order to determine whether the treatment revisions have impacted the course of change in the expected direction. This process may be repeated any number of times. In-treatment retestings also can provide information relevant to the decision of when to terminate treatment.

The goal of monitoring is to determine whether treatment is "on track" with the progress that is expected at a given point in time. When and how often one might assess the patient is dependent on a few factors. The first is the instrumentation. Many instruments are designed to assess the patient's status at the time of testing. Items on these measures are generally worded in the present tense (e.g., "I feel tense and nervous," "I feel that my family loves and cares about me"). Changes from one day to the next on the constructs measured by the instrument should be reflected in the test results.

Other instruments, however, ask the patient to indicate if a variable of interest has been present, or how much or to what extent it has occurred during a specific time period in the past. The items usually are asked in the context of something like "During the past month, how often have you . . ." or "During the past week, to

what extent has ...” Readministration of a measure containing interval-of-time-specific items or subsets of items should be undertaken only after a period of time equivalent to or longer than the time interval to be considered in responding to the items has past. For example, an instrument which asks the patient to consider how much certain symptoms have been problematic during the past seven days should not be readministered for at least seven days. The responses elicited during a readministration that occurs less than seven days after the first administration would include the patient’s consideration of his or her status during the previously considered time period. This may make interpretation of the change of symptom status (if any) from the first to the second assessment difficult if not impossible.

Methods to determine whether clinically significant change has occurred from one point in time to another have been developed and can be used for treatment monitoring purposes. These are discussed in the outcomes assessment section of this chapter below. However, for monitoring purposes, another approach to evaluating therapeutic change may be superior. This approach may be referred to as the “glide path” approach, with the term referring to the narrow descent course or path that airplanes must follow when landing. Deviation from the flight glide path requires corrections in the plane’s speed, altitude, and/or attitude in order to return to the glide path and a safe landing.

R.L. Kane (personal communication, July 22, 1996) has indicated that just as a pilot has the instrumentation to alert him or her about the plane’s position on the glide path, the clinician may use psychological assessment instruments to track how well the patient is following the glide path of treatment. The glide path in this case represents expected improvement over time in one or more measurable areas of functioning (e.g., symptom severity, social role functioning, occupational performance). The expectations would be based on objective data obtained from similar patients at various points during their treatment and would allow for minor deviations from the path. The end of the glide path is one or more specific goals that are part of the treatment plan. Thus, “arrival” at the end of the glide path signifies the attainment of specific treatment goals.

4.18.7 PSYCHOLOGICAL ASSESSMENT AS A THERAPEUTIC INTERVENTION

The use of psychological assessment as a means of therapeutic intervention in and of

itself has received more than passing attention during the past few years. “Therapeutic assessment” with the MMPI-2 has received particular attention primarily through the work of Finn and his associates (Finn, 1996a, 1996b; Finn & Martin, in press; Finn & Tonsager, 1992). Finn’s approach appears to be applicable with instruments or batteries of instruments that provide multidimensional information relevant to the concerns of patients seeking answers to questions related to their mental health status. The approach espoused by Finn thus will be presented here as a model for deriving direct therapeutic benefits from the psychological assessment experience.

4.18.7.1 What Is Therapeutic Assessment?

In discussing the use of the MMPI-2 as a therapeutic intervention, Finn (1996a) describes an assessment procedure whose goal is to “gather accurate information about clients ... and then use this information to help clients understand themselves and make positive changes in their lives” (p. 3). Elaborating on this procedure and extending it to the use of any test, Finn and Martin (in press) describe therapeutic assessment as

collaborative, interpersonal, focused, time limited, and flexible. It is ... very interactive and requires the greatest of clinical skills in a challenging role for the clinician. It is unsurpassed in a respectfulness for clients: collaborating with them to address *their* concerns (around which the work revolves), acknowledging them as experts on themselves and recognizing their contributions as essential, and providing to them usable answers to their questions in a therapeutic manner.

The ultimate goal of therapeutic assessment is to provide an experience for the client that will allow him/her to take steps toward greater psychological health and a more fulfilling life. This is done by recognizing the client’s characteristic ways of being, understanding in a meaningful, idiographic way the problems the client faces, providing a safe environment for the client to explore change, and providing the opportunity for the client to experience new ways of being in a supportive environment.

Simply stated, therapeutic assessment may be considered an approach to the assessment of mental health patients in which the patient is not only the primary provider of information needed to answer questions, but also is actively involved in formulating the questions that are to be answered by the assessment. Feedback regarding the results of the assessment is provided to the patient and is considered a

primary, and possibly the principal element of the assessment process. Thus, the patient becomes a partner in the assessment process; as a result, therapeutic and other benefits accrue.

The reader should note that in this section, the term “therapeutic assessment” is used to denote the specific approach advocated by Finn and his colleagues for using the psychological assessment process as an opportunity for therapeutic intervention. It should not be confused with the more general term “therapeutic psychological assessment” as it has been employed throughout this chapter; “therapeutic assessment” is but one aspect of therapeutic psychological assessment.

4.18.7.2 The Impetus for Therapeutic Assessment

To say that clinical psychologists performing psychological assessments in mental health settings traditionally have never shared much of their findings with their patients is probably not an overstatement. A common scenario throughout many mental health settings was that a patient being treated by a psychologist was evaluated by the latter, or a patient was referred to the psychologist by another mental health professional for assessment only. In the first instance, the degree to which the psychologist might directly share the results of the often lengthy and expensive evaluation would vary. Generally, a detailed review of the findings would be a rarity. In the latter instance, the patient would be evaluated, a report of the results dictated, and a copy of the report sent back to the referring clinician. In either instance, the purpose of the assessment probably would be to answer questions posed by the treating clinician. Unfortunately, the patient and his or her concerns as they related to the psychological assessment were typically of only secondary consideration, if any.

Fortunately, recent occurrences have begun to change the way in which assessment information is used. Consequently, the degree to which the patient is involved in the assessment process is changing. One reason for this is the relatively recent revision of the ethical standards of the APA (1992). This revision included a mandate for psychologists to provide feedback to clients whom they test. According to ethical standard 2.09:

Unless the nature of the relationship is clearly explained to the person being assessed in advance and precludes provision of an explanation of results (such as in some organizational consulting, pre-employment or security screenings, and forensic evaluations), psychologists ensure that an explanation of the results is provided using language that is

reasonably understandable to the person assessed or to another legally authorized person on behalf of the client. Regardless of whether the scoring and interpretation are done by the psychologist, by assistants, or by automated or other outside services, psychologists take reasonable steps to ensure that appropriate explanations of results are given. (p. 8)

Many clinicians and other psychologists involved in assessment activities (e.g., counseling psychologists, neuropsychologists) have had to modify their practice routine to accommodate this requirement. Some view this requirement as resulting in an improvement in the quality of their services; others likely see it as nothing more than an inconvenience which, in the era of managed care and limited access to treatment, further limits the amount of time they have to work with a patient. However, most would agree that the patient has benefited from the required feedback.

Finn and Tonsager (1992) identified other factors that may have contributed to the recent interest in providing patients with assessment feedback. One is another external influence, that is, the recognition of the patient’s right to see their medical and psychiatric health care records. However, they also point to several clinically and research-based findings and impressions that suggest that therapeutic assessment enhances patient care through the facilitation of patient–therapist rapport, cooperation during the assessment process, positive feelings about the process and the clinician, improvement in mental health status, and feelings of being understood by another. In addition, Finn and Tonsager refer to Finn and Butcher’s (1991) summary of potential benefits that may accrue from providing test results feedback. The listed benefits, based on clinical experience, include increased feelings of self-esteem and hope, reduced symptomatology and feelings of isolation, increased understanding and self-awareness, and increased motivation to seek or be more actively involved in mental health treatment. Finally, Finn and Martin (in press) note that the therapeutic assessment process can lead to increased feelings of mastery and control and decreased feelings of alienation. At the same time, it can serve as a model for relationships that can result in mutual respect and the patient being seen for who he or she is.

4.18.7.3 The Therapeutic Assessment Process

Finn (1996a) has outlined a three-step procedure for therapeutic assessment using the MMPI-2. As indicated above, it should

work equally well with other multidimensional instruments that one might select. Finn describes this procedure as one to be used in those situations in which the patient is seen only for assessment (i.e., the patient is not to be treated later by the assessing clinician). From the present author's standpoint, the procedures are equally applicable for use by clinicians who test patients whom they later treat. With these points in mind, the three-step procedure is summarized below.

4.18.7.3.1 Step 1: The initial interview

According to Finn (1996a), the initial interview with the patient serves multiple purposes. It provides an opportunity to build rapport, or to increase rapport if a patient-therapist relationship already exists. The assessment task is presented as a collaborative one, and the patient is given the opportunity to identify questions that he or she would like answered using the assessment data. Background information related to the patient-identified questions is subsequently gathered. Any reservations about participating in the therapeutic assessment process (e.g., confidentiality, previous negative experiences with assessment) are dealt with in order to facilitate maximal involvement in the process.

After responding to the patient's concerns, Finn (1996a) recommends that the clinician restate the questions posed earlier by the patient. This ensures the accuracy of what the patient would like to have addressed by the assessment. The patient also is encouraged to ask questions of the clinician, thus reinforcing the collaborative context or atmosphere that the clinician is trying to establish. Step 1 is completed as the instrumentation and its administration, as well as the responsibilities and expectations of each party, are clearly defined and the particulars of the process (e.g., date and time of assessment, date and time of the feedback session, clinician fees) are discussed and agreed upon.

4.18.7.3.2 Step 2: Preparing for the feedback session

Upon completion of the administration and scoring of the instrumentation used during the assessment, the clinician first outlines all results obtained from the assessment, including those not directly related to the patient's previously stated questions. Finn (1996a) presents a well-organized outline for the types of information that the trained user can extract from MMPI-2 data. These include response consistency, test-taking attitude, distress and disturbance, major symptoms, underlying personality, behavior in

relationships, implications for treatment, diagnostic impression, and recommendations. Unfortunately, clinicians who do not or cannot use the MMPI-2 or other well-researched, multidimensional instruments will not have the same amount or type of data available to them. (This should not preclude them from identifying the types of valid and useful information that can be derived from the instruments and organizing it into a usable form for presentation to the patient.) This is followed by a determination of how to present the results to the patient. This can be guided by the clinician asking himself or herself the following questions:

- (i) How do the (test) findings relate to the client's goals?
- (ii) What are the most important findings of the (tests administered)?
- (iii) To what extent is the client likely to already know about and agree with the (test) findings?
- (iv) How much new information is the client likely to be able to integrate in the feedback session?
- (v) What is likely to happen if the client becomes overwhelmed or is presented with findings that are greatly discrepant from his/her current self-concept? (p. 34)

As a final point in this step, Finn (1996a) indicates that the clinician must determine what is the best way to present the information to the patient so that he or she can accept and integrate the information while maintaining his or her sense of identity and self-esteem. This also is a time when the clinician can identify information that he or she may not wish to reveal to the patient because it is not important to answering the patient's questions; doing so may negatively affect the collaborative relationship. In addition, the clinician may want to prepare for presenting those aspects of feedback that he or she feels will be most problematic for him or her (i.e., the clinician) by role-playing with a colleague.

4.18.7.3.3 Step 3: The feedback session

As Finn (1996a) states: "The overriding goal of feedback sessions is to have a therapeutic interaction with clients" (p. 44). Thus, the initial tasks of the feedback session are focused on setting the stage for this type of encounter. This is accomplished by allaying any anxiety the patient may have about the session, reaffirming the collaborative relationship, and familiarizing him or her with the presentation of the test results (e.g., explaining the profile sheet upon which the results are graphed, discussing the normative group to which he or she will be compared, providing an explanation of standard scores).

When the session preparation is completed, the clinician begins providing feedback to the patient (Finn, 1996a). This, of course, is centered on answering the questions posed by the patient during Step 1. Beginning with a positive finding from the assessment, the clinician proceeds to first address those questions that the patient is most likely to accept. He or she then carefully moves to the findings that are more likely to be anxiety-arousing for the patient and/or challenge his or her self-concept. A key element to this step is to have the patient verify the accuracy of each finding and provide a real-life example of the interpretation that is offered. Alternately, one should ask the patient to modify the interpretation to make it more in line with how he or she sees themselves and their situation. Finn (1996a) provides specific suggestions about how to deal with a rejection of a finding, the final suggestion being to allow the client to disagree with but not totally dismiss the finding. This leaves the door open for representing the finding at another time when the patient is more open to accepting it.

Finn (1996a) recommends that the clinician should end the session by responding to any additional questions the patient may have; confirming that the patient has accurately understood the information that was presented; giving permission for the patient to contact the clinician should further questions arise; and (in the assessment-only arrangement) termination of the relationship. Throughout the session, the clinician maintains a supportive stance with regard to any affective reactions to the findings.

4.18.7.3.4 Additional steps

Finn and Martin (in press) indicate two additional steps that may be added to the therapeutic assessment process. The purpose of the first additional step, referred to as an "assessment intervention session" essentially is to clarify initial test findings through the administration of additional instruments. For example, Finn and Martin explain how MMPI-2 findings can be further fleshed out through a nonstandard administration of an instrument such as the Thematic Apperception Test (TAT). Here, the clinician controls the patient's interpretation in order to draw out information relevant to the patient's questions. Also, solutions to problems elicited by the TAT cards are suggested to the patient.

The other additional step discussed by Finn and Martin (in press) is the provision of a written report of the findings to the patient. In addition to summarizing both the test results and the answers to the patient's questions, it also

attempts to elicit feedback and reactions from the patient about the assessment.

The reader should note that the preceding summary presents only the key technical aspects of the therapeutic assessment procedures espoused by Finn and his associates. Much of the clinical/dynamic aspect of this approach has not been addressed because of the focus of this chapter. Those interested in incorporating the process into their clinical practice are encouraged to read Finn (1996a).

4.18.7.4 Empirical Support for Therapeutic Assessment

Noting the lack of direct empirical support for the therapeutic effects of sharing test results with patients, Finn and Tonsager (1992) investigated the benefits of providing feedback to university counseling center clients regarding their MMPI-2 results. A total of 32 subjects underwent therapeutic assessment and feedback procedures similar to those described above while on the counseling center's waiting list. Another 28 subjects were recruited from the same waiting list to serve as a control group. There were no significant differences between the two groups on any important demographic or examiner contact-interval variables.

Instead of receiving feedback, Finn and Tonsager's (1992) control group received non-therapeutic attention from the examiner. However, they were administered the same dependent measures as the feedback group at the same time as the experimental group received feedback. They also were administered the same dependent measures as the experimental group two weeks later (i.e., two weeks after the experimental group received the feedback) in order to determine if there were differences between the two groups on those dependent measures. These measures included a self-esteem questionnaire, a symptom checklist (i.e., the SCL-90-R), a measure of private and public self-consciousness, and a questionnaire assessing the subjects' subjective impressions of the feedback session. (Note that the control group was administered only that portion of the feedback assessment questionnaire that was relevant to them.)

The results of Finn and Tonsager's (1992) study indicated that compared to the control group, the feedback group demonstrated significantly less distress at the two-week post-feedback follow up, and significantly higher levels of self-esteem and hope at both the time of feedback and the two-week post-feedback follow up. In other findings, feelings about

the feedback sessions were positively and significantly correlated with changes in self-esteem from testing to feedback, both from feedback to follow up and from testing to follow up among those who were administered the MMPI-2. In addition, the change in level of distress from feedback to follow up correlated significantly with private self-consciousness (i.e., the tendency to focus on the internal aspects of oneself) but not with public self-consciousness.

4.18.8 PSYCHOLOGICAL ASSESSMENT AS A TOOL FOR OUTCOMES MANAGEMENT

The 1990s have witnessed a positively accelerating growth curve reflecting the level of interest in and development of behavioral health care outcomes programs. Cagney and Woods (1994) attribute this to four major factors. First, behavioral health care purchasers are asking for information regarding the value of the services they buy. Second, there is an increasing number of purchasers who are requiring a demonstration of patient improvement and satisfaction. Third, MCOs need data that demonstrate that their providers render efficient and effective services. And fourth, outcomes information will be needed for the "quality report cards" that MCOs anticipate they will be required to provide in the future. In short, fueled by soaring health care costs, there has been an increasing need for providers to demonstrate that what they do is effective. And all of this has occurred within the context of the continuous quality improvement (CQI) movement, in which there have been similar trends in the level of interest and growth.

As this author has noted previously, the interest in and necessity for outcomes measurement in the era of managed care and accountability provides a unique opportunity for clinical psychologists to use their training and skills in assessment (Maruish, 1994). However, the extent to which the clinical psychologist becomes a key and successful contributor to an organization's outcomes initiative (whatever that might be) will depend on his or her understanding of what "outcomes" and their measurement and applications are all about.

4.18.8.1 What Are Outcomes?

Before discussing outcomes, it is important to have a clear understanding of what is meant by the term. Experience has shown that the meaning varies according to whom one may speak.

Donabedian (1985) has identified three dimensions of quality of care. "Structure" refers to the organization providing the care. It includes aspects such as how the organization is "organized," the physical facilities and equipment, and the number and professional qualifications of its staff. "Process" refers to the specific types of services that are provided to a given patient (or group of patients) during a specific episode of care. These might include various types of tests and assessments (e.g., psychological tests, lab tests, magnetic resonance imaging), therapeutic interventions (e.g., group psychotherapy, medication), and discharge planning activities. Treatment complications (e.g., drug reactions) are also included here. "Outcomes", on the other hand, refers to the results of the specific treatment that was rendered.

The outcomes, or results, of treatment should not refer to change in only a single aspect of functioning. Treatment may impact various facets of a patient's life. Stewart and Ware (1992) have identified five broad aspects of general health status: physical health, mental health, social functioning, role functioning, and general health perception. Treatment may affect these aspects of health in different ways, depending on the disease or disorder being treated and the effectiveness of the treatment. Some specific aspects of functioning related to these five areas of general health status that are commonly measured include: feeling of well-being, psychological symptom status, use of alcohol and other drugs, functioning on the job or at school, marital/family relationships, utilization of health care services, and ability to cope.

In considering the various types of outcomes that might be assessed in behavioral health care settings, a substantial number of clinicians probably would identify symptomatic change in psychological status as being the most important. Nevertheless, however important change in symptom status may have been in the past, clinical psychologists and other behavioral health care providers have come to realize that changes in many of the other aspects of functioning identified by Stewart and Ware (1992) are equally important indicators of treatment effectiveness. As Sederer et al. (1996) have noted:

Outcome for patients, families, employers, and payers is not simply confined to symptomatic change. Equally important to those affected by the care rendered is the patient's capacity to function within a family, community, or work environment or to exist independently, without undue burden to the family and social welfare

system. Also important is the patient's ability to show improvement in any concurrent medical and psychiatric disorder ... Finally, not only do patients seek symptomatic improvement, but they want to experience a subjective sense of health and well being. (p. 2)

A much broader perspective is offered in Faulker and Gray's *The 1995 behavioral outcomes and guidelines sourcebook* (Migdal, Youngs, & Bengen-Seltzer, 1995):

Outcomes measures are being redefined from a vague "is the patient doing better?" to more specific questions, such as, "Does treatment work in ways that are measurably valuable to the patient in terms of daily functioning level and satisfaction, to the payor in terms of value for each dollar spent, to the managed care organization charged with administering the purchaser's dollars, and to the clinician charged with demonstrating value for hours spent?" (p. 1)

Thus, "outcomes" holds a different meaning for each of the different parties who have a stake in behavioral health care delivery. What is measured generally depends on the purposes for which outcomes assessment is undertaken. As will be shown, these vary greatly.

4.18.8.2 Outcomes Assessment: Measurement, Monitoring, and Management

Just as it is important to be clear about what is meant by outcomes, it is equally important to clarify the three general purposes for which outcomes assessment may be employed. The first is outcomes measurement. This involves nothing more than pre- and post-treatment assessment of one or more variables to determine the amount of change that has occurred (if any) in these variables as a result of therapeutic intervention.

A more useful approach is that of outcomes monitoring. This refers to "the use of periodic assessment of treatment outcomes to permit inferences about what has produced change" (Dorwart, 1996, p. 46). Like treatment progress monitoring used for treatment planning purposes, outcomes monitoring involves the tracking of changes in the status of one or more outcomes variables at multiple points in time. Assuming a baseline assessment at the beginning of treatment, reassessment may occur one or more times during the course of treatment (e.g., weekly, monthly), at the time of termination, and/or during one or more periods of post-termination follow up. Whereas treatment progress monitoring is used to determine how much the patient is on or off the expected course of improvement, outcomes monitoring focuses

on revealing aspects about the therapeutic process that seem to affect change.

The third and most useful purpose of outcomes assessment is that of outcomes management. Dorwart (1996) defines outcomes management as "the use of monitoring information in the management of patients to improve both the clinical and administrative processes for delivering care" (pp. 46–47). In outcomes management, information is used to improve the quality of services offered to the patient population(s) served by the provider, not to any one patient. Information gained through the assessment of patients can provide the organization with indications of what works best with whom and under what set of circumstances, thus helping to improve the quality of services for all patients. In essence, outcomes management can serve as a tool for those organizations with an interest in implementing a CQI initiative.

4.18.8.3 The Benefits of Outcomes Assessment

The implementation of any type of outcomes assessment initiative within an organization does not come without effort from and cost to the organization. However, if it is implemented properly, all interested parties, that is, patients, clinicians, provider organizations, payers, and the health care industry as a whole, should find a substantial yield from the outlay of time and money. Cagney and Woods (1994) identify several benefits to patients, including enhanced health and quality of life, improved health care quality, and effective use of the dollars paid into benefits plans. For providers, the outcomes data can result in improved clinical skills, information related to the quality of care provided and local practice standards, increased profitability, and decreased concerns over possible litigation.

Outside of the clinical context, benefits also can accrue to payers and MCOs. Cagney and Woods (1994) see the potential payer benefits as including healthier workers, improved health care quality and worker productivity, and reduced or contained health care costs. As for MCOs, the benefits include increased profits, information that can help shape the practice patterns of their providers, and decisions that are based on quality of care.

4.18.8.4 The Therapeutic Use of Outcomes Assessment

The foregoing overview of outcomes assessment provides the background necessary for discussing the use of psychological outcomes assessment data in day-to-day clinical practice.

Whereas the focus of the above review was centered on both the individual patient and patient populations, it now will narrow to the use of outcomes assessment primarily in service to the individual patient. The reader interested in issues related to large, organization-wide outcomes studies conducted for outcomes management purposes (as defined above) is encouraged to seek other sources of information that specifically address that topic (see, for example, Migdail, Youngs, & Bengden-Seltzer, 1995; Newman, 1994).

There is no one system or approach to the assessment of treatment outcomes for an individual patient that is appropriate for all providers of behavioral health care services. Because of the specific type of outcomes one is interested in, the reasons for assessing them, and the manner in which they may impact the decisions made by the patient, payer and clinician, any successful and useful outcomes assessment approach must be customized. Customization should reflect the needs of the primary beneficiary of the information gained from the assessment (i.e., patient, payer, or provider), with consideration of the secondary stakeholders in the therapeutic endeavor. Ideally, the identified primary beneficiary would be the patient. Although this is not always the case, it would appear that only rarely would the patient not benefit, at least indirectly, from the gathering of outcomes data.

Following are considerations and recommendations for the development and implementation of an outcomes assessment initiative by behavioral health care providers. Although space limitations do not allow a comprehensive review of all issues and solutions, the information that follows can be useful to clinical psychologists and others with similar training wishing to begin to incorporate outcomes assessment into their standard therapeutic routine.

4.18.8.4.1 Purpose of the outcomes assessment

There are numerous reasons for assessing outcomes. For example, in a recent survey of 73 behavioral health care organizations, various reasons were identified by the participants as to why they had conducted outcomes studies (Pallak, 1994). Among the several indicated, the top five reasons (in descending order) were to: evaluate outcomes for patients, evaluate provider effectiveness, evaluate integrated treatment programs, manage individual patients, and support sales and marketing efforts. However, from the clinician's standpoint, a couple of purposes are worth noting.

In addition to monitoring the course of progress during treatment (see above), clinicians may employ outcomes assessment to obtain a direct measure of how much patient improvement has occurred as the result of the course of treatment intervention. Here, the findings are of more benefit to the clinician than to the patient himself because a pre- and post-treatment approach to the assessment is utilized. The information will not lead to any change in the patient providing the information, but the feedback it provides to the clinician could assist him in the treatment of other patients later on.

Another common reason for outcomes assessment is to demonstrate the patient's need for therapeutic services beyond that which is typically covered by the patient's medical and behavioral health care benefits. When assessment is conducted for this reason, the patient and clinician are only secondary beneficiaries of the outcomes data. As will be shown below, the type of information that a third party payer requires for authorization of extended benefits may not be the most relevant or useful to the patient or the clinician.

4.18.8.4.2 What to measure

The aspects or dimensions of patient functioning that are measured as part of outcomes assessment will depend on the purpose for which the assessment is being conducted. As discussed earlier, probably the most commonly measured variable is that of symptomatology or psychological/mental health status. After all, disturbance or disruption in this dimension is probably the most common reason why people seek behavioral health care services in the first place. However, there are other reasons for seeking help, including difficulties in coping with various types of life transitions (e.g., a new job, recent marriage or divorce, other changes in the work or home environment), inability to deal with the behavior of others (e.g., spouse, children), general dissatisfaction with life, or perhaps other less common reasons. Additional assessment of related variables therefore may be necessary, or may even take precedence over the assessment of symptoms or other mental health indicators.

In the vast majority of the cases seen for behavioral health care services, the assessment of the patient's overall level of psychological distress or disturbance will yield the most singularly useful information, regardless of whether it is used for outcomes measurement, outcomes monitoring, outcomes management, or to meet the requirements of third-party

payers for authorization of additional benefits. Indices such as the Positive Symptom Total (PST) or Global Severity Index (GSI) that are part of the SA-45 or BSI can provide this type of information efficiently and economically.

For some patients, measures of one or more specific psychological disorders or symptom clusters are at least as important if not more important than overall symptom or mental health status. Here, if interest is in only one disorder or symptom cluster (e.g., depression), one may choose to measure only that particular set of symptoms using an instrument designed specifically for that purpose (e.g., use of the BDI with depressed patients). For those interested in assessing the outcomes of treatment relative to multiple psychological dimensions, the administration of more than one disorder-specific instrument or a single, multidimensional instrument which assesses all or most of the dimensions of interest would be required. Again, instruments such as the SA-45 or the BSI can provide a quick, broad assessment of multiple symptom domains. Although much lengthier, other multiscale instruments, such as the MMPI-2 or the PAI, permit a more detailed assessment of several disorders or symptom domains using one inventory.

In many cases, the assessment of mental health status is adequate for outcomes assessment purposes. There are other instances in which changes in psychological distress or disturbance either provide only a partial indication of the degree to which therapeutic intervention has been successful, are not of interest to the patient or a third-party payer, are unrelated to the reason why the patient sought services in the first place, or are otherwise inadequate or unacceptable as measures of improvement in the patient's condition. One may find that for some patients, improved functioning on the job, at school, or with family or friends is much more relevant and important than symptom reduction. For other patients, improvement in their quality of life or feeling of well-being is more meaningful.

It is not always a simple matter to determine exactly what should be measured. However, careful consideration of the following questions should greatly facilitate the decision.

(i) *Why did the patient seek services?* People pursue treatment for many reasons. The patient's stated reason for seeking therapeutic assistance may be the first clue in determining what is important to measure.

(ii) *What did the patient hope to gain from treatment?* The patient's stated goals for the treatment he or she is about to receive may be a primary consideration in the selection of outcomes to be assessed.

(iii) *What are the patient's criteria for the successful completion of the current therapeutic episode?* The patient's goals for treatment may provide only a broad target for the therapeutic intervention. Having the patient identify exactly what will have to happen to consider treatment successful and no longer needed will help in specifying the most important constructs and/or behaviors to assess.

(iv) *What are the clinician's criteria for the successful completion of the current therapeutic episode?* What the patient identifies as being important to accomplish during treatment may reflect a lack of insight into his or her problems, or it might not otherwise concur with what the clinician's experience would indicate.

(v) *What are the criteria of significant third parties for the successful completion of the current therapeutic episode?* From a strict treatment perspective, this should be given the least amount of consideration. From a more realistic perspective, one cannot overlook the expectations and limitations that one or more third parties have for the treatment that is rendered. The expectations and limitations set by the patient's behavioral health care plan, the guidelines of the organization in which the clinician practices, and possibly other external forces may significantly play into the decision about when to terminate treatment.

(vi) *What, if any, are the outcomes initiatives within the provider organization?* One cannot ignore any outcomes programs that have been initiated by the organization in which the therapeutic services are delivered. Regardless of the problems and goals of the individual patient, organization-wide studies of effectiveness may dictate the gathering of specific types of outcomes data from patients who have received services.

Note that the selection of the variables to be assessed may address more than one of the above issues. Ideally, this is what should happen. However, one needs to take care that the gathering of outcomes data does not become too burdensome. As a general rule, the more outcomes data one attempts to gather from a given patient or collateral, the less likely one is to obtain any data at all. The key is to identify the point at which the amount of data that can be obtained from a patient and/or collaterals, and the ease with which it can be gathered, is optimized.

4.18.8.4.3 How to measure

Once the decision concerning what to measure has been made, one must then decide how this should be measured. In many cases, the most important data will be that obtained

directly from the patient through the use of self-report instruments. Underlying this assertion are the assumptions that valid and reliable instrumentation, appropriate to the needs of the patient, is available to the clinician; the patient can read at the level required by the instruments; and the patient is motivated to respond honestly to the questions asked. If this is not the case, other options are available.

Other types of data-gathering tools may be substituted for self-report measures. Rating scales completed by the clinician or other members of the treatment staff may provide information that is as useful as that elicited directly from the patient. In those cases in which the patient is severely disturbed, unable to give valid and reliable answers (e.g., younger children), unable to read, or is an otherwise inappropriate candidate for a self-report measure, clinical rating scales can substitute as useful means of gathering data. Related to these instruments are parent-completed inventories for child and adolescent patients. These are particularly useful in obtaining information about the child or teen's behavior that might not otherwise be known.

Collateral rating instruments can also be used to gather information in addition to that obtained from self-report measures. When used in this manner, these instruments provide a mechanism by which the clinician and other treatment staff can contribute data to the outcomes assessment endeavor. This not only results in the clinician or provider organization having more information upon which to evaluate the outcomes of therapeutic intervention, it also gives the clinician an opportunity to ensure that the perspective of the treatment provider is considered in the evaluation of the effects of the treatment given.

Another potential source of outcomes information is administrative data. In many of the larger provider organizations, this information is easily retrieved through their management information systems (MISs). Data related to the patient's diagnosis, dose and regimen of medication, physical findings, course of treatment, and other types of data typically stored in these systems can be useful to those evaluating the outcomes of therapeutic intervention.

4.18.8.4.4 When to measure

There are no hard and fast rules, guidelines, or accepted conventions related to when outcomes should be assessed. The common practice is to assess the patient at least at treatment initiation and treatment termination/discharge. Obviously, at the time of treatment initiation, the clinician should obtain a baseline measure of

whatever variables will be measured at the termination. At the minimum, this allows for "outcomes measurement" as described above. As has been discussed, additional assessment of the patient on the variables of interest can take place at other points in time, that is, at other times during the course of treatment and upon post-discharge follow up.

Many would argue that postdischarge/post-termination follow-up assessment provides the best or most important indication of the outcomes of therapeutic intervention. Two types of comparisons may be made on follow-up. The first is a comparison of the patient's status on the variables of interest at the time of treatment initiation, or at the time of discharge or termination, to that of the patient at some point after treatment has ended. Either way, this follow-up data will provide an indication of the more lasting effects of the intervention. Generally, the variables of interest for this type of comparison include such things as symptom presence and intensity, feeling of well-being, frequency of substance use, and social and role functioning.

The second type of post-treatment investigation involves looking at the frequency at which some aspect(s) of the patient's life circumstances, behavior or functioning occurred during a given period prior to treatment, compared to that which occurred during an equivalent period of time immediately preceding the post-discharge assessment. This approach is commonly used in determining the cost-offset benefits of treatment. For example, the number of times a patient has been seen in an emergency room for psychiatric problems during the three-month period preceding the initiation of outpatient treatment can be compared to the number of emergency room visits during the three-month period preceding the postdischarge follow-up assessment. Not only can this provide an indication of the degree to which treatment has helped the patient deal with his problems, it can also demonstrate how much medical expenses have been reduced through the patient's decreased use of costly emergency room services.

In general, post-discharge outcomes assessment probably should take place no sooner than a month after treatment has ended. When feasible, one probably should wait three to six months to assess the variables. This should provide a more valid indication of the lasting effects of treatment. Assessments being conducted to determine the frequency with which some behavior or event occurs (as may be needed to determine cost-offset benefits) can be accomplished no sooner than the reference time interval used in the baseline assessment. Thus,

suppose that the patient reports 10 emergency room visits during the three-month period prior to treatment. If one wants to know if the patient's emergency room visits have decreased after treatment, the assessment cannot take place any earlier than three months after treatment termination.

4.18.8.4.5 How to analyze outcomes data

There are two general approaches to the analysis of treatment outcomes data. The first is to determine whether changes in patient scores on outcomes measures are statistically significant. The other is to establish whether these changes are clinically significant. Use of standard tests of statistical significance is important in the analysis of group or population change data. Clinical significance is more relevant to change in the individual patient. As this chapter is focused on the individual patient, this section will center on matters related to determining clinically significant change as the result of treatment.

The issue of clinical significance has received a great deal of attention in psychotherapy research during the past several years. This is at least partially owing to the work of Jacobson and his colleagues (Jacobson, Follette, & Revenstorf, 1984, 1986; Jacobson & Truax, 1991) and others (e.g., Christensen & Mendoza, 1986; Speer, 1992; Wampold & Jenson, 1986). Their work came at a time when researchers began to recognize that traditional statistical comparisons do not reveal a great deal about the efficacy of therapy. In discussing the topic, Jacobson and Truax broadly define the clinical significance of a treatment as

its ability to meet standards of efficacy set by consumers, clinicians, and researchers. While there is little consensus in the field regarding what these standards should be, various criteria have been suggested: a high percentage of clients improving . . . ; a level of change that is recognizable by peers and significant others . . . ; an elimination of the presenting problem . . . ; normative levels of functioning at the end of therapy . . . ; high end-state functioning at the end of therapy . . . ; or changes that significantly reduce one's risk for various health problems. (p. 12)

From their perspective, Jacobson and his colleagues (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991) felt that clinically significant change could be conceptualized in one of three ways. Thus, for clinically significant change to have occurred, the measured level of functioning following the therapeutic episode would either:

(i) fall outside the range of the dysfunctional population by at least two standard deviations away from the mean of that population, in the direction of functionality;

(ii) fall within two standard deviations of the mean for the normal or functional population; or

(iii) be closer to the mean of the functional population than to that of the dysfunctional population.

Jacobson and Truax viewed the third option as being the least arbitrary and provided different recommendations for determining cut-offs for clinically significant change, depending upon the availability of normative data. Lambert (1994) demonstrated how the third option could be modified to allow for the inclusion of more than one categorization of dysfunction (e.g., mild, moderate, severe). This assumes, of course, that the necessary normative data needed to separate the gradations of dysfunction are available.

At the same time, these same investigators noted the importance of considering the change in the measured variables of interest from pre- to post-treatment, in addition to the patient's functional status at the end of therapy. To this end, Jacobson et al. (1984) proposed the concomitant use of a reliable change (RC) index to determine whether change is clinically significant. This index, modified on the recommendation of Christensen and Mendoza (1986), is nothing more than the pretest score minus the posttest score divided by the standard error of the difference of the two scores, expressed as:

$$RC = (x_2 - x_1)/S_{\text{diff}}$$

where x_1 is the pretest score, x_2 is the post-test score, and S_{diff} is the standard error of the difference. The standard error of the difference is computed as:

$$S_{\text{diff}} = \sqrt{2 (\text{SEM})^2}$$

where SEM is the standard error of measurement for a functional group (e.g., normals, nonpatients) on the instrument. If the RC index is greater than 1.96, the change in scores is not likely to be due to chance ($p < 0.05$), but rather to reflect real change.

Speer (1992) recommended a different approach when regression to the mean has been demonstrated to contribute to the improvement in scores from pre- to post-test. The alternate approach, based on the combined work of Nunnally (1967) and Edwards, Yarvis, Mueller, Zingale, and Wagman (1978), involves developing a confidence interval of ± 2 SEMs around the estimated true pretest score. A post-test

score falling outside of this confidence interval is considered significantly different from the initial pretest score at $p < 0.05$. Using this approach, more change is needed to show clinically significant improvement than to show clinically significant deterioration. Note that the criterion for determining whether regression to the mean is occurring is met when a negative correlation is found to exist between the pretreatment score and amount of change that has taken place. This implies the evaluation of group data, and for this reason this empirical criterion may not be of use for the individual patient unless the latter is a member of a sample for which test results are available.

Lambert (1994) proposes a modified recommendation for the dual criteria for clinically significant change (that is, RC greater than 1.96 and movement of the patient's score from the dysfunctional group's distribution to the functional group's distribution) such that movement from one degree of dysfunction to a lesser degree would also meet one of the two criteria for clinically significant change. In an example, Lambert illustrated that normative data for the Global Severity Index (GSI) found in the SCL-90-R literature can be used to empirically define four levels of symptom intensity: asymptomatic, mildly symptomatic, moderately symptomatic, and severely symptomatic. Assuming an RC of 1.96 or greater, clinically significant change can be said to have occurred if a patient's GSI score moves from severely to moderately or mildly symptomatic, or to asymptomatic; from moderately to mildly symptomatic, or to asymptomatic; or from mildly symptomatic to asymptomatic. Although this criterion is less stringent than having to move from being symptomatic (regardless of the severity) to asymptomatic, it still provides information that is quite useful for clinical decision making.

4.18.9 FUTURE DIRECTIONS

The ways in which clinical psychologists have conducted the types of psychological assessment described in this chapter have undergone dramatic changes during the 1990s. This should come as no surprise to anyone who spends a few minutes a day skimming the newspaper or watching television. The health care revolution started gaining momentum at the beginning of the 1990s and has not since slowed down. And there are no indications that it will subside in the foreseeable future. There was no real reason to think that behavioral health care would be spared from being a target of the revolution, and there is no good reason why it should have been spared. The behavioral health care

industry certainly had contributed its share of waste, inefficiency, and lack of accountability to the problems that led to the revolution. Now, like other areas of health care, it is forced to "clean up its act." Some consumers of mental health or chemical dependency services have benefited from the revolution, others have not. In any case, the way in which health care is delivered and financed has changed, and clinical psychologists and other behavioral health care professionals must adapt to survive in the market.

Some of those involved in the delivery of psychological assessment services may wonder (with some fear and trepidation) where the revolution is leading the behavioral health care industry and, in particular, how their ability to practice will be affected. At the same time, others are eagerly awaiting the inevitable advances in technology and other resources that come with the passage of time. What will occur is open to speculation. However, close observation of the practice of psychological assessment and the various industries that support it (particularly the forms of therapeutic assessment described in this chapter) has led this author to arrive at some predictions as to where the field of therapeutic psychological assessment is headed and the implications these have for clinicians, provider organizations, and patients. What follows in this section are the most important of these predictions. Also included are what this author feels are the needs that must be met if psychological assessment is to continue to be a valued contributor to the delivery of efficient, cost-effective behavioral health care.

4.18.9.1 What the Industry Is Moving Away From?

One way of discussing what the field is moving toward is to first talk about what it is moving away from. In the case of therapeutic psychological assessment, two trends are becoming quite clear. First, starting at the beginning of this last decade of the twentieth century, the use of (and reimbursement for) psychological assessment has gradually been curtailed. This particularly has been the case with regard to indiscriminate assessment involving the administration of lengthy and expensive batteries of psychological tests. Payers began to demand evidence that the knowledge gained from the administration of these instruments contributed to the delivery of cost-effective, efficient care to mental health and substance abuse patients. There are no indications that this trend will stop.

Second, assessment has begun to move away from the use of lengthy, multidimensional objective instruments (e.g., the MMPI) or time-consuming projective techniques (e.g., Rorschach) that previously represented the standard of practice. When assessment is authorized now, it usually involves the use of inexpensive yet well-validated, problem-oriented instruments. This reflects modern behavioral health care's time-limited, problem-oriented approach to treatment. The clinician can no longer afford to spend a great deal of time in assessment activities when the patient has only a limited number of payer-authorized sessions with him or her. Thus, both now and in the foreseeable future, brief instruments will be used for problem identification or clarification, progress monitoring, and/or outcomes assessment.

4.18.9.2 Trends in Instrumentation

The move toward the use of brief, problem-oriented instruments for therapeutic psychological assessment purposes has just been identified. Another trend in the selection of instrumentation is the increasing use of public domain tests, questionnaires, rating scales, and other types of measurement tools. Previously, these free-use instruments were not developed with the rigor that is usually applied in the development of psychometrically sound instruments by commercial test publishers. Consequently, they typically lacked the validity and reliability data that are necessary to judge their psychometric integrity.

Recently, however, there has been a significant improvement in the quality and documentation of the public domain and other "for-free" tests that are available for therapeutic psychological assessment. Instruments such as the SF-36/SF-12 and HSQ/HSQ-12 health measures are good examples of such tools. These and instruments such as the Behavior and Symptom Identification Scale (BASIS-32; Eisen, Grob, & Klein, 1986) and the Outcome Questionnaire (OQ-45.1; Lambert, Lunnen, Umphress, Hansen, & Burlingame, 1994) have undergone psychometric scrutiny and have gained widespread acceptance. Although copyrighted, they may be used for a nominal one-time or annual licensing fee; thus, they generally are treated much like public domain assessment tools. One can expect that other quality, useful instruments will be made available for use at little or no cost in the future.

As for the types of instrumentation that will be needed and developed, one can probably expect some changes. Accompanying the increasing focus on outcomes assessment is a

recognition by payers and patients that changes in several areas of functioning are at least as important as changes in level of symptom severity when evaluating the effectiveness of the treatment. For example, employers are interested in the patient's ability to resume the functions of his or her job, while family members may be concerned with the patient's ability to resume their role as spouse or parent. Increasingly, measurement of the patient's functioning in areas other than psychological/mental status has come to be included as part of behavioral health care outcomes systems. Probably the most visible indication of this is the incorporation of the SF-36 or HSQ into various behavioral health care studies, and the fact that two major psychological test publishers offer HSQ products in their catalogs of clinical products. One will likely see other public domain and commercially available nonsymptom-oriented instruments, especially those emphasizing social and occupational role functioning, in increasing numbers over the next several years.

Other types of instrumentation will also become prominent. These will include measures of variables that support the outcomes and other therapeutic assessment initiatives undertaken by provider organizations. What one organization or provider feels is important, or what it is told is important for reimbursement or other purposes, will dictate what is measured. Instrumentation will also include measures that will be useful for the prediction of outcomes for individuals seeking psychotherapeutic services from those organizations.

4.18.9.3 Trends in Data Use and Storage

There are two areas of application in which the valuable data obtained from therapeutic psychological assessment have heretofore been overlooked or underutilized. Indications are that this will change for both in the future. One area for which assessment data has potential application is that of clinical decision-making. This of course pertains only to the use of outcomes assessment data. Generally, data gathered solely for the purpose of outcomes assessment is used for just that: the assessment of the results of treatment. This is particularly the case in large, formal outcomes management programs. As has been discussed earlier in this chapter, data gathered at the beginning of treatment can be used immediately for treatment planning purposes while also serving as baseline data that can be compared to discharge data later on.

The other potential area of data application is in the development of local, regional, and

national databases of therapeutic assessment data. Patient data gathered by various providers, organizations or programs within organizations, at one or more points during the therapeutic episode, can be pooled and used for various purposes. These databases can then serve as the bases for two highly beneficial (and probably profitable) endeavors. The first is the generation of sets of normative data for various populations delineated along any number of parameters. Norms for any number of instruments or health care variables could be generated "on demand" and continuously updated to reflect trends in behavioral health care. This author is aware of one large, national behavioral health care system where such a database already exists. He also is aware of efforts at establishing cross-organizational databases of this kind.

The second benefit afforded by the information contained in these databases is that of predictive modeling. For example, the behavioral health care organization mentioned above has taken advantage of the organizational data available to it to investigate the relationships between a number of treatment, demographic and other variables and the outcomes of treatment. Subjecting the large data sets available to it to sophisticated statistical analyses has allowed this organization to determine those types of patients requiring special care or attention in order to achieve desired outcomes at the time of treatment termination. Predictive modeling can also be used for identifying variables related to other aspects of patient care, such as patient satisfaction with the care received. The possibilities for the use of data in this manner are enormous.

4.18.9.4 Trends in the Application of Technology

Clinical psychologists have not been shy when it has come to taking advantage of the technological advances that have been achieved since the late 1970s. This is no more evident than in the extent to which the personal computer and the vast array of psychological assessment software have been incorporated into their delivery of clinical services. Automated administration, scoring, and interpretation and reporting of the results of nearly all major objective tests are currently available to the clinician through PC-based software. In addition, the availability of affordable desktop optical scanners allows the clinician to maintain the portability of the assessment instruments while retaining the scoring and interpreting power of the computer for processing the test data.

To speculate on how technology will be advanced in the service of therapeutic psychological assessment in the future is a risky business. As has been witnessed since the late 1970s, much can happen quickly. There are, however, three areas of technologic or technology-dependent advances on the horizon to which clinical psychologists should have access in the not too distant future. The first is the availability of online administration, scoring, and interpretation and reporting of tests via the Internet. In fact, an Internet version of the SA-45 is being beta tested at the time of writing (1997). To this author's knowledge, this represents the first use of the Internet for psychological assessment purposes. It is anticipated that the Internet version of the SA-45 will be commercially available in the very near future, and it will be quickly followed by the availability of Internet versions of other assessment instruments.

The second advance is actually a technology that has been around for a while but has undergone improvements, that is, the fax-back technology that is being used for scoring and reporting of objective, paper-and-pencil tests. Essentially, the fax machine replaces the optical scanner as a means of data entry. The electronic data is entered directly into the test publisher's computer for processing and report generation. However, instead of generating a hard-copy report of results at the processing site, the report is transmitted in electronic form and sent back to the clinician's fax machine within minutes of processing. At that point, the report is printed out just like any other fax transmission. Currently, this technology is being used on a somewhat limited basis. This is partially owing to the degree to which test publishers are making this form of automated scoring and reporting available to their customers. However, in the relatively near future, one should see more tests being offered to clinicians in this manner, particularly as the technology continues to improve.

The third area of technologic advancement has more to do with the application of technology than the development of new technology. L. E. Beutler and O. B. Williams (personal communication, January 15, 1996) have taken Beutler and Clarkin's (1990) Systematic Treatment Selection (STS) model of prescriptive treatment assignment and have developed specifications for software for automating the matching of treatments, therapists, and patients. The capability of subsequently tracking patients during the course of treatment is also included in these specifications. Originally entitled *STS for Windows*, this software is now under development through a behavioral

health care publishing and consulting company. Driving the STS system is patient assessment data related to six variables: subjective distress, functional severity, problem complexity, potential for therapeutic resistance, coping style, and social support. Each of these variables may be assessed through either commercially available self-report instruments or clinician rating scales developed specifically for use with STS.

When fully developed, the STS should serve as the standard for in-office treatment–patient–therapist matching and patient-tracking software. According to the developers (L. E. Beutler & O. B. Williams, personal communication, January 15, 1996), the software will include numerous features, the most important of which will be: a comprehensive treatment planning report with up-to-date references to relevant research articles and treatment manuals; automatic entry of each patient's data into a growing database that is used for treatment planning and prediction; the ability to predict the amount of symptom reduction from a specific course of therapy; a report profiling the patient's symptom status over time; a report indicating individual clinician's ability to treat specific types of symptomatology; and the ability to incorporate case notes into the patient's electronic file. The major benefits of the system include the ability to: use different assessment means (self-report or clinician rating) to obtain the information needed to drive the system; develop treatment recommendations based on information that is optimal for the patient; easily monitor patient progress on a glide path developed from the treatment of similar patients and adjust the therapy plan (if necessary) on a timely basis; and determine a clinician's therapeutic strengths and weaknesses, thus permitting the most effective patient–therapist match.

All in all, when fully developed, the STS software will combine the knowledge and expertise of a leader in the field of psychotherapeutic research with state-of-the-art technology, thus yielding a powerful decision-making behavioral health care tool. One can be assured that similar products are likely to follow once the benefits of the STS software become widely known.

4.18.10 SUMMARY

The health care revolution has brought mixed blessings to those in the behavioral health care professions. It has resulted in limitations for reimbursement for services rendered and has forced many to change the way they practice their profession. At the same time, it has led to revelations about the cost savings benefits that

can accrue from the treatment of mental health and substance use disorders. This has been the bright spot in an otherwise bleak picture for some behavioral health care professionals. For clinical psychologists, the picture appears to be somewhat different. They now have additional opportunities to contribute to the positive aspects of the revolution and to gain from the “new order” it has imposed. By virtue of their training in psychological assessment and through the application of appropriate instrumentation, they are uniquely qualified to support or otherwise facilitate multiple aspects of the therapeutic process. It is the clinical psychologist's contributions to aspects of “therapeutic psychological assessment” that this chapter has sought to identify and address in some detail.

Earlier, this author identified some of the types of psychological assessment instruments that are commonly used in the service of therapeutic endeavors. These included both brief and lengthy (multidimensional) symptom measures, as well as measures of general health status, quality of life, and patient satisfaction with the services received. Also identified were different sets of general criteria that can be applied when selecting instruments for use in therapeutic settings. The main intent of this chapter, however, was to present a detailed discussion of the various therapeutic uses of psychological assessment.

Generally, psychological assessment can assist the clinician in three important clinical activities: clinical decision-making, treatment itself (when used as a specific therapeutic technique), and treatment outcomes assessment. Regarding the first of these activities, three important clinical decision-making functions can be facilitated by psychological assessment: screening, treatment planning, and treatment monitoring. The first of these can be served by the use of “down and dirty” instruments to identify, within a high degree of certainty, the likelihood of the presence (or absence) of a particular condition or characteristic. Here, the diagnostic efficiency of the instrument used (as indicated by the PPP and NPP) is of great importance. Through their ability to identify and clarify problems as well as other important treatment-relevant patient characteristics, psychological assessment instruments can also be of great assistance in planning treatment. In addition, treatment monitoring, or the regular determination of the patient's progress throughout the course of treatment, can be served well by the application of psychological assessment instruments.

Secondly, assessment may be used as part of a therapeutic technique. In what Finn terms

“therapeutic assessment,” situations in which patients are evaluated via psychological testing are used as opportunities for the process itself to serve as a form of therapeutic intervention. This is accomplished through involving the patient as an active participant in the assessment process, not just as the object of the assessment.

Thirdly, psychological assessment can be employed as the primary mechanism by which the outcomes or results of treatment can be measured. However, the use of assessment for this purpose is not a cut-and-dried matter. As discussed, there are issues, pertaining to what to measure, how to measure, and when to measure, that require considerable thought prior to undertaking any standard (or even nonstandard) plan to assess outcomes. Guidelines for resolving these issues are presented, as is information pertaining to how to determine whether the measured outcomes of treatment are indeed “significant.”

In the final section of the chapter, this author shares some thoughts about where psychological assessment is probably heading in the future. No radical revelations are presented since no signs really point in that direction. What is foreseen is the appearance of more quality assessment instruments that will remain in the public domain, and greater application of communications technology, fax and the Internet, in particular, as assessment delivery, scoring and reporting mechanisms. Also predicted is the application of tomorrow’s computer technology to available assessment data for optimized treatment–patient–therapist matching. The innovative proposals of Beutler and Williams in this regard seem to represent the state-of-the-art thinking at this time.

There is no doubt that the practice of psychological assessment has been dealt a blow within recent years. However, as this chapter hopefully has shown, clinical psychologists have the skills to take this powerful tool, apply it in ways that will benefit those suffering from mental health and substance abuse problems, and demonstrate its benefits and their skills to patients and payers. Whether they will be successful in this demonstration will be determined in the near future. In the meantime, advances will continue to be made that will facilitate their work and improve its quality.

4.18.11 REFERENCES

- American Psychological Association (1992). *Ethical principles of psychologists*. Washington, DC: Author.
- American Psychological Association (1996). *The costs of failing to provide appropriate mental health care*. Washington, DC: Author.
- Andrews, G., Peters, L., & Teesson, M. (1994). *The measurement of consumer outcomes in mental health*. Canberra, Australia: Australian Government Publishing Service.
- Appelbaum, S. A. (1990). The relationship between assessment and psychotherapy. *Journal of Personality Assessment*, 54, 791–801.
- Attkisson, C. C., & Zwick, R. (1982). The Client Satisfaction Questionnaire: Psychometric properties and correlations with service utilization and psychotherapy outcome. *Evaluation and Program Planning*, 6, 233–237.
- Baldessarini, R. J., Finkelstein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry*, 40, 569–573.
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guilford Press.
- Beutler, L. E., & Clarkin, J. (1990). *Systematic treatment selection: Toward targeted therapeutic interventions*. New York: Brunner/Mazel.
- Beutler, L. E., Wakefield, P., & Williams, R. E. (1994). Use of psychological tests/instruments for treatment planning. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 55–74). Hillsdale, NJ: Erlbaum.
- Beutler, L. E., & Williams, O. B. (1995). Computer applications for the selection of optimal psychosocial therapeutic interventions. *Behavioral health care Tomorrow*, 4, 66–68.
- Burlingame, G. M., Lambert, M. J., Reisinger, C. W., Neff, W. M., & Mosier, J. (1995). Pragmatics of tracking mental health outcomes in a managed care setting. *Journal of Mental Health Administration*, 22, 226–236.
- Butcher, J. N. (1990). *The MMPI-2 in psychological treatment*. New York: Oxford University Press.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
- Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. (1989). *Development and use of the MMPI-2 content scales*. Minneapolis, MN: University of Minnesota Press.
- Cagney, T., & Woods, D. R. (1994). Why focus on outcomes data? *Behavioral health care Tomorrow*, 3, 65–67.
- Center for Disease Control and Prevention (1994, May 27). Quality of life as a new public health measure: Behavioral risk factor surveillance system. *Morbidity and Mortality Weekly Report*, 43, 375–380.
- Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index [Letter to the editor]. *Behavior Therapy*, 17, 305–308.
- Ciarlo, J. A., Brown, T. R., Edwards, D. W., Kiresuk, T. J., & Newman, F. L. (1986). *Assessing mental health treatment outcomes measurement techniques*. DHHS Pub. No. (ADM)86–1301. Washington, DC: US Government Printing Office.
- Commission on Chronic Illness (1987). *Chronic illness in the United States: 1*. Cambridge, MA: Commonwealth Fund, Harvard University Press.
- Derogatis, L. R. (1983). *SCL-90-R: Administration, scoring and procedures manual-II*. Baltimore: Clinical Psychometric Research.
- Derogatis, L. R. (1992). *BSI: Administration, scoring and procedures manual-II*. Baltimore: Clinical Psychometric Research.
- Derogatis, L. R., & DellaPietra, L. (1994). Psychological tests in screening for psychiatric disorder. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 22–54). Hillsdale, NJ: Erlbaum.

- Derogatis, L. R., Lipman, R. S., & Covi, L. (1973). SCL-90: An outpatient psychiatric rating scale—preliminary report. *Psychopharmacology Bulletin*, 9, 13–27.
- Dickey, B., & Wagenaar, H. (1996). Evaluating health status. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 55–60). Baltimore: Williams & Wilkins.
- Donabedian, A. (1985). *Explorations in quality assessment and monitoring: Vol. III. The methods and findings of quality assessment monitoring: An illustrative analysis*. Ann Arbor, MI: Health Administration Press.
- Dorwart, R. A. (1996). Outcomes management strategies in mental health: Applications and implications for clinical practice. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 45–54). Baltimore: Williams & Wilkins.
- Dowd, E. T., Milne, C. R., & Wise, S. L. (1991). The therapeutic Reactance Scale: A measure of psychological reactance. *Journal of Counseling and Development*, 69, 541–545.
- Edwards, D. W., Yarvis, R. M., Mueller, D. P., Zingale, H. C., & Wagman, W. J. (1978). Test-taking and the stability of adjustment scales: Can we assess patient deterioration? *Evaluation Quarterly*, 2, 275–292.
- Eisen, S. V., Grob, M. C., & Klein, A. A. (1986). BASIS: The development of a self-report measure for psychiatric inpatient evaluation. *The Psychiatric Hospital*, 17, 165–171.
- Elwood, R. W. (1993). Psychological tests and clinical discrimination: Beginning to address the base rate problem. *Clinical Psychology Review*, 13, 409–419.
- Ficken, J. (1995). New directions for psychological testing. *Behavioral Health Management*, 20, 12–14.
- Finn, S. E. (1996a). *Manual for using the MMPI-2 as a therapeutic intervention*. Minneapolis, MN: University of Minnesota Press.
- Finn, S. E. (1996b). Assessment feedback integrating MMPI-2 and Rorschach findings. *Journal of Personality Assessment*, 67, 543–557.
- Finn, S. E., & Butcher, J. N. (1991). Clinical objective personality assessment. In M. Hersen, A. E. Kazdin, & A. S. Bellack (Eds.), *The clinical psychology handbook* (2nd ed., pp. 362–373). New York: Pergamon.
- Finn, S. E., & Martin, H. (in press). Therapeutic assessment with the MMPI-2 in managed health care. In J. N. Butcher (Ed.), *Objective personality assessment in managed health care: A practitioner's guide*. Minneapolis, MN: University of Minnesota Press.
- Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment*, 4, 278–287.
- Friedman, R., Sobel, D., Myers, P., Caudill, M., & Benson, H. (1995). Behavioral medicine, clinical health psychology, and cost offset. *Health Psychology*, 14, 509–518.
- Gough, H. G., McClosky, H., & Meehl, P. E. (1951). A personality scale for dominance. *Journal of Abnormal and Social Psychology*, 46, 360–366.
- Gough, H. G., McClosky, H., & Meehl, P. E. (1952). A personality scale for social responsibility. *Journal of Abnormal and Social Psychology*, 47, 73–80.
- Greene, R. L. (1991). *The MMPI-2/MMPI: An interpretive manual*. Boston: Allyn & Bacon.
- Greene, R. L., & Clopton, J. R. (1994). Minnesota Multiphasic Personality Inventory—2. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 137–159). Hillsdale, NJ: Erlbaum.
- Greenfield, T. K., & Attkisson, C. C. (1989). Progress toward a multifactorial service satisfaction scale for evaluating primary care and mental health services. *Evaluation and Program Planning*, 12, 271–278.
- Hathaway, S. R., & McKinley, J. C. (1951). *MMPI manual*. New York: The Psychological Corporation.
- Health Outcomes Institute (1993). *Health Status Questionnaire 2.0 manual*. Bloomington, MN: Author.
- Hsiao, J. K., Bartko, J. J., & Potter, W. Z. (1989). Diagnosing diagnoses: Receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry*, 46, 664–667.
- Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, 15, 336–352.
- Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1986). Toward a standard definition of clinically significant change [Letter to the editor]. *Behavior Therapy*, 17, 309–311.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19.
- Jahoda, M. (1958). *Current concepts of mental health*. New York: Basic Books.
- Lambert, M. J. (1994). Use of psychological tests for outcome assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 75–97). Hillsdale, NJ: Erlbaum.
- Lambert, M. J., Lunnen, K., Umphress, V., Hansen, N. B., & Burlingame, G. M. (1994). *Administration and scoring manual for the Outcome Questionnaire (OQ-45.1)*. Salt Lake City, UT: IHC Center for Behavioral health care Efficacy.
- Larsen, D. L., Attkisson, C. C., Hargreaves, W. A., & Nguyen, T. D. (1979). Assessment of client/patient satisfaction: Development of a general scale. *Evaluation and Program Planning*, 2, 197–207.
- LeVois, M., Nguyen, T. D., & Attkisson, C. C. (1981). Artifact in client satisfaction assessment: Experience in community mental health settings. *Evaluation and Program Planning*, 4, 139–150.
- Maruish, M. (1990, Fall). Psychological assessment: What will its role be in the future? *Assessment Applications*, 7–8.
- Maruish, M. E. (1994). Introduction. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 3–21). Hillsdale, NJ: Erlbaum.
- Megargee, E. I., & Spielberger, C. D. (1992). Reflections on fifty years of personality assessment and future directions for the field. In E. I. Megargee & C. D. Spielberger (Eds.), *Personality assessment in America* (pp. 170–190). Hillsdale, NJ: Erlbaum.
- Mental Health Weekly*. (1996a, April 8). Future targets behavioral health field's quest for survival, 1–2.
- Mental Health Weekly*. (1996b, April 8). Leaders predict integration of MH primary care by 2000, 1–6.
- Mental Health Weekly*. (1996c, April 29). Anxiety disorders screening day occurs this week, 7.
- Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283–298.
- Migdal, K. J., Youngs, M. T., & Bengen-Seltzer, B. (Eds.) (1995). *The 1995 behavioral outcomes & guidelines sourcebook*. New York: Faulkner & Gray.
- Millon, T. (1994). *MCMI-III manual*. Minneapolis, MN: National Computer Systems.
- Moreland, K. L. (1996). How psychological testing can reinstate its value in an era of cost containment. *Behavioral health care Tomorrow*, 5, 59–61.
- Morey, L. C. (1991). *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.
- Morey, L. C., & Henry, W. (1994). Personality Assessment Inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 185–216). Hillsdale, NJ: Erlbaum.
- Newman, F. L. (1994). Selection of design and statistical

- procedures for progress and outcome assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 111–134). Hillsdale, NJ: Erlbaum.
- Newman, F. L., & Ciarlo, J. A. (1994). Criteria for selecting psychological instruments for treatment outcome assessment. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 98–110). Hillsdale, NJ: Erlbaum.
- Nguyen, T. D., Attkisson, C. C., & Stegner, B. L. (1983). Assessment of patient satisfaction: Development and refinement of a service evaluation questionnaire. *Evaluation and Program Planning*, 6, 299–313.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
- Olfson, M., & Pincus, H. A. (1994a). Outpatient psychotherapy in the United States, I: Volume, costs, and user characteristics. *American Journal of Psychiatry*, 151, 1281–1288.
- Olfson, M., & Pincus, H. A. (1994b). Outpatient psychotherapy in the United States, II: Patterns of utilization. *American Journal of Psychiatry*, 151, 1289.
- Oss, M. E. (1996). Managed behavioral health care: A look at the numbers. *Behavioral Health Management*, 16, 16–17.
- Pallak, M. S. (1994). National outcomes management survey: Summary report. *Behavioral health care Tomorrow*, 3, 63–69.
- Phelps, R. (1996, February). Preliminary practitioner survey results enhance APA's understanding of health care environment. *Practitioner Focus*, 9, 5.
- Psychotherapy Finances* (1995, January). Fee, practice and managed care survey, 21(1), Issue 249.
- Radosevich, D., & Pruitt, M. (1996). *Twelve-item Health Status Questionnaire (HSQ-12) version 2.0 user's guide*. Bloomington, MN: Health Outcomes Institute.
- Radosevich, D. M., Wetzler, H., & Wilson, S. M. (1994). *Health Status Questionnaire (HSQ) 2.0: Scoring comparisons and reference data*. Bloomington, MN: Health Outcomes Institute.
- Schlosser, B. (1995). The ecology of assessment: A "patient-centric" perspective. *Behavioral health care Tomorrow*, 4, 66–68.
- Sederer, L. I., Dickey, B., & Hermann, R. C. (1996). The imperative of outcomes assessment in psychiatry. In L. I. Sederer & B. Dickey (Eds.), *Outcomes assessment in clinical practice* (pp. 1–7). Baltimore: Williams & Wilkins.
- Simmons, J. W., Avant, W. S., Demski, J., & Parisher, D. (1988). Determining successful pain clinic treatment through validation of cost effectiveness. *Spine*, 13, 34.
- Sipkoff, M. Z. (1995, August). Behavioral health treatment reduces medical costs: Treatment of mental disorders and substance abuse problems increases productivity in the workplace. *Open Minds*, 12.
- Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology*, 60, 402–408.
- Spielberger, C. D. (1983). *Manual of the State-Trait Anxiety Inventory: STAI (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.
- Stewart, A. L., & Ware, J. E., Jr. (1992). *Measuring functioning and well-being*. Durham, NC: Duke University Press.
- Strain, J. J., Lyons, J. S., Hammer, J. S., Fahs, M., Lebovits, A., Paddison, P. L., Snyder, S., Strauss, E., Burton, R., & Nuber, G. (1991). Cost offset from a psychiatric consultation-liaison intervention with elderly hip fracture patients. *American Journal of Psychiatry*, 148, 1044–1049.
- Strategic Advantage, Inc. (1996). *Symptom Assessment-45 Questionnaire manual*. Minneapolis, MN: Author.
- Substance Abuse Funding News*, (1995, December 22). Brown resigns drug post. 7.
- VandenBos, G. R., DeLeon, P. H., & Belar, C. D. (1993). How many practitioners are needed? It's too early to know! *Professional Psychology: Research and Practice*, 22, 441–448.
- Vermillion, J., & Pfeiffer, S. (1993). Treatment outcome and continuous quality improvement: Two aspects of program evaluation. *Psychiatric Hospital*, 24, 9–14.
- Wampold, B. E., & Jenson, W. R. (1986). Clinical significance revisited [Letter to the editor]. *Behavior Therapy*, 17, 302–305.
- Ware, J. E., Kosinski, M., & Keller, S. D. (1995). *SF-12: How to Score the SF-12 Physical and Mental summary scales* (2nd ed.). Boston: New England Medical Center, The Health Institute.
- Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-Item Short Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, 30, 473–483.
- Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health Survey manual and interpretation guide*. Boston: New England Medical Center, The Health Institute.
- Werthman, M. J. (1995). A managed care approach to psychological testing. *Behavioral Health Management*, 15, 15–17.
- World Health Organization (1948). Constitution. In *Basic Documents*. Geneva, Switzerland: Author.

4.19

Forensic Assessment

DAVID FAUST

University of Rhode Island, Kingston, RI, USA

4.19.1 INTRODUCTION	563
4.19.2 SOME KEY ISSUES AND CONSIDERATIONS IN PERSONAL INJURY CASE	564
4.19.2.1 <i>Four Central Issues</i>	564
4.19.2.2 <i>The Four Elements and the Forensic Examiner's Task</i>	565
4.19.2.3 <i>Treating Professionals Called into Court</i>	568
4.19.2.3.1 <i>Maintaining the treatment alliance as primary</i>	568
4.19.2.3.2 <i>Release of assessment or treatment records</i>	569
4.19.3 ADMISSIBILITY	570
4.19.4 ASSESSMENT METHODS	573
4.19.4.1 <i>Informed Consent</i>	573
4.19.4.2 <i>Clarifying Referral Questions and Deciding Whether they are Appropriate and Within the Clinician's Expertise</i>	573
4.19.4.3 <i>Access to Information</i>	574
4.19.4.4 <i>Design of Assessment Procedures: Some Do's and Don't Do's</i>	575
4.19.4.4.1 <i>Use the best available methods</i>	575
4.19.4.4.2 <i>Obtaining adequate information</i>	578
4.19.4.4.3 <i>Conduct technically proficient evaluations</i>	580
4.19.4.4.4 <i>Give adequate consideration to alternatives</i>	582
4.19.4.5 <i>Interpretive Strategies</i>	584
4.19.4.6 <i>Preparing Reports</i>	587
4.19.5 LAWYERS' STRATEGIES AND TACTICS	588
4.19.5.1 <i>Credentials</i>	589
4.19.5.2 <i>Bias</i>	589
4.19.5.3 <i>Manner of Conducting the Examination</i>	590
4.19.5.4 <i>Erroneous or Questionable Conclusions</i>	590
4.19.5.5 <i>Scientific Status</i>	590
4.19.6 DEPOSITIONS AND TRIAL TESTIMONY	591
4.19.6.1 <i>A Sampling of Deposition Topics</i>	591
4.19.6.2 <i>Some Suggestions for Depositions</i>	594
4.19.6.3 <i>Trial Testimony</i>	596
4.19.7 REFERENCES	598

4.19.1 INTRODUCTION

Psychologists interface with the courts in various contexts and situations. A developmental psychologist may describe a 3 year old's recognition of risk in a product liability case in which a child was injured playing with a toy. A clinical psychologist may testify about the more

suitable parent in a custody dispute in family court, or may opine about a defendant's mental state when he committed murder in a criminal case. A neuropsychologist may give evidence about cognitive or brain functions in a personal injury case. Other areas of testimony might include such diverse topics as the trustworthiness of eye witnesses reports, the appropriateness of a

statistical analysis purportedly linking a toxin with a disorder, test bias in screening job applicants, and the characteristics of individuals who delay a considerable period of time before reporting traumatic events.

Such qualitative variation in courtroom activities precludes coverage of all these topics in one chapter and necessitates an alternative strategy. Much of the material to follow will have a generic quality, that is, many of the points that will be raised relate to a broad array of courtroom work. However, such generic discussion will be anchored by, or mainly directed towards, the civil arena, with an emphasis on distress claims (e.g., post-traumatic stress disorder) and brain injury claims. These topics provide suitable illustration of many of the more general points that will be raised. Additionally, distress and brain injury claims are very active areas in psychology and law, and hence of likely interest to a broad audience. Although assessment is the organizing theme of the chapter, this activity is not entirely separable from the context of courtroom opinions and other collateral issues (e.g., depositions), and thus the material that follows will not be rigidly limited to evaluative methodology. Finally, readers who are mainly interested in a highly specialized area (e.g., child custody, or multiple personality in criminal cases) may supplement the material presented here with additional sources.

The chapter will begin with coverage of basic issues that frame psychologists' courtroom involvement and that are often the focus of legal proceedings, such as the core elements of civil cases and whether psychologists participate as treating professionals or as retained experts. I will then discuss, in turn, the admissibility of psychological evidence; methods for conducting and improving courtroom evaluations; and case strategies and tactics, including lawyers' main areas of attack, depositions, direct testimony, and cross-examination.

The emphasis of this chapter is far more practical than philosophical or theoretical: the main intent is to provide materials that will enhance the quality of mental health professionals' legal work. Other psychologists may take a different view than my own on the level or type of scientific backing that is appropriate for legal work, although my views on this matter are considerably more moderate than some readers might suppose (e.g., I simply maintain that psychological testimony should normally rest on fundamentally sound scientific foundations [see Faust, 1993]). Further, I believe that psychology is making substantial strides, and that there are an increasing number of areas in which testimony is likely to be of considerable scientific and practical merit in the courtroom

(see the upcoming revision of Faust, Ziskin, & Hiers, 1991; Faust, Ziskin, Hiers, & Miller, in press). Whatever differences there might be in viewpoint about the needed level of scientific backing for legal testimony, which, of course, is ultimately a matter for the courts to decide, presumably all psychologists would agree that there is merit in increasing the quality of courtroom work: it is with that spirit and intent that this chapter is written.

Some psychologists are hardened veterans of courtroom work and others are just beginning to consider courtroom involvement. This chapter is directed mainly towards those at an intermediate or beginning level. Further, some of the points I raise are based on my experience with legal work, particularly as a consultant. I am not assuming that my experience with legal cases is necessarily representative of such cases overall, and I am not using my observations to try to reach scientific conclusions or generalizations, but rather to provide what I hope will be practical guidance.

The guidelines and suggestions for the practice of forensic psychological assessment detailed in this chapter emerge from, and are directed towards, the USA court system, and their generalization and application to other court systems vary. Nevertheless, many of the main points likely will apply to forensic work in other countries, in particular the emphasis on methods and strategies that increase the likelihood of accurate conclusions.

4.19.2 SOME KEY ISSUES AND CONSIDERATIONS IN PERSONAL INJURY CASE

4.19.2.1 Four Central Issues

Personal injury cases, which involve civil (and not criminal) law, address situations in which it is claimed that one or more individuals have not carried out some duty owed to one or more other individuals, or have not exercised reasonable care. For example, Smith may have been negligent in failing to shovel the driveway, and Jones may have slipped on the ice and suffered a broken arm. Many personal injury cases involve lay individuals, not professionals. Thus, the legal issues frequently relate to responsibilities of the general citizenry, such as exercising reasonable care when driving a car or shooting a gun.

Individuals are held to obligations and standards commensurate with the role in which they are functioning and the assumed levels of knowledge and responsibility that are expected or required. For example, in a medical malpractice case, a neurosurgeon will almost surely be held to a much higher standard in recognizing

a brain tumor than a clinical psychologist with a general practice. At the same time, both the neurosurgeon's and the psychologist's responsibilities are dictated by their roles as professionals. Once they take on a patient's case, they are not only obligated to the duties that bind the ordinary citizen, but in addition the responsibilities involved in the professional care of a patient. In contrast, and obviously, ordinary citizens are not obligated to properly diagnose and treat individuals for medical or psychological disorders. The citizen may be responsible for the broken arm resulting from the fall on his unshoveled driveway, but not for interpreting the x ray or setting the cast properly.

In most legal cases, in order for a plaintiff to prevail, or at least to make a monetary recovery, four elements, and all four elements, must be proven (this also applies to malpractice cases). First, there must be a duty owed. For example, we might owe a duty to maintain our premises free of foreseeable hazards, but not to maintain our neighbor's premises. Second, it must be proven that there was a breach of duty, or that reasonable care was not exercised. For example, if we do not shovel our sidewalk, we may well have breached our duty. But if an earthquake caused our sidewalk to crack and someone fell and hurt himself before the rumbling stopped and action could be taken, the plaintiff likely will not prevail.

Third, the breach of duty must be a cause of the harm. If the person fell, but it was because her feet got tangled in the dog's leash and had nothing to do with the tools left scattered on the stoop of the shop, the business owner is not liable. Depending on the circumstances, issues, and other possible factors, the potential agents in question may need to be the main cause of the harm, a substantial contributor, or may only have to have made any type of meaningful (versus trivial) contribution. Damages are sometimes awarded in relation to the relative contribution of the cause at issue. Thus, if the event is assumed to account for 10% of the outcome, one multiplies the award by .10. In other cases, if a factor made any significant contribution, the defendant is responsible for the total damage that ensued. For example, if the defendant's careless driving caused the accident, then it makes no difference that most of the plaintiff's injuries would have been avoided had she worn a seat belt.

The fourth and final issue is damages. Assuming the first three criteria are met and the defendant is hence found to be blameworthy (i.e., liable), plaintiffs are to be compensated for any damages stemming from the event. A duty that was owed may have been breached and caused an undesired event, but if that event

resulted in no damages or minimal damages, there will be little or no award (and the case, in fact, will probably go nowhere in the courts). For example, a therapist may have engaged in an egregious action, such as throwing a suicidal patient out of treatment the moment he discovered the patient's insurance had run out. However, if no harm resulted, for example, the patient did not make a suicide attempt and through some odd quirk ended up benefiting from the termination, there are no damages for which the patient is to be compensated, although the courts or a professional board could still sanction the professional.

4.19.2.2 The Four Elements and the Forensic Examiner's Task

Depending on the type of civil case in which a psychologist is involved as an expert and the particulars, any or all of the four elements discussed may be pertinent to the professional's task. In malpractice cases, for example, the professional may address each of the elements. There may be debate about whether: (a) a brief or even casual contact with a troubled person created a professional duty or relationship, or about the nature of that duty; (b) whether the psychologist has followed reasonable care in assessing suicidal ideation or has met a certain standard of care; (c) whether failure to meet some practice standard caused the harm the person suffered; and (d) about the type and extent of harm that occurred. However, in many other cases, mental health professionals do not touch on at least the first two issues, that is, whether or what duty was owed and whether it was breached. For example, a clinical psychologist usually does not testify about proper driving practices.

The issue of causation may or may not be central to the psychologist's courtroom testimony. In some cases, another expert will testify about cause. For example, a neurologist may indicate that the plaintiff's cognitive disturbance is due to brain damage suffered in the accident at issue. In such a case, a neuropsychologist might primarily address the issue of damages, or the cognitive and behavioral impairments that are present. There may be little need for the neuropsychologist to link the damages to brain injury because another professional, the neurologist in this case, has already made the causal connection. The plaintiff's attorney may prefer to leave the issue of cause in the neurologist's hands, or the neuropsychologist may be restricted from providing causal testimony. In some cases there may be multiple causal elements to be considered, one or more of which the

psychologist will address. For example, the issue of whether reckless driving caused the accident can be separate from an issue such as whether the accident caused brain damage. The first question about cause involves the assignment of blame for the event, and the second the link between the event and its possible consequences. The jury could decide that an improper left-hand turn caused the accident, and that the accident caused a chronic neck condition but not a brain injury.

In other cases, however, if the mental health professional does not testify to cause, there may be no other expert who can. For example, in a malpractice case, it will likely be up to the psychiatrist or psychologist to connect a therapist's breach of duty with the plaintiff's problems, or at least those problems linked to the event (such as depression, increased anxiety, or decreased responsiveness to therapeutic intervention). If there are plausible alternative causes for the problems (e.g., substance abuse or other emotional traumas), and if the professional herself cannot determine whether there is a causal link, how can one expect the jury to make a judgment about a matter the expert cannot decide? A decent cross-examiner can thus quickly show that with all the professional's presumed expertise, she cannot say there is a link between the event and adverse functioning. This may be devastating to the plaintiff's case. In some instances, if the expert's inability to address cause comes to light prior to trial (e.g., during a deposition that is taken as part of the discovery process), this may lead the judge to place tight restrictions on the psychologist's trial testimony, or even to exclude the expert or dismiss the entire case. For example, in a toxin case in which brain damage is being claimed, if the attorney can produce no expert to help with this core element—causality—it may not be possible to prove it legally and the case may be over.

The need to establish cause also illustrates some of the differences that can arise across clinical and courtroom situations. In clinical practice, a determination of cause or etiology may make little difference because it may not alter treatment. In the courtroom, the specific cause, whatever the treatment implications, may make all of the difference. A psychologist may fail to recognize how an issue that can be of such minimal relevance in a clinical situation, and, therefore, sometimes bypassed or set aside during the course of an evaluation, might have huge repercussions at deposition or trial. This is also why, in the legal arena, it is often not enough to identify a condition as present. It may be true that Smith has brain damage, but if it was not the car accident but a toxic exposure in a different place and circumstance that caused it, the driver of the other vehicle is not responsible.

It follows that careful consideration of alternate causes, including procurement of relevant documents and information, is often crucial to the psychologist's legal work.

Psychologists usually should have a basic familiarity with key legal issues in a case before they undertake, or, preferably, before they decide whether to undertake, a courtroom evaluation, and should not assume that the lawyer will convey this information. The psychologist is typically better off knowing in advance whether cause is at issue and whether the lawyer does (or does not) want the psychologist to address this matter; and whether the standard applicable to the case is a substantial versus primary contributor. Such foreknowledge may allow psychologists to identify circumstances in which they are being asked to do something that is unreasonable or beyond their expertise, or may allow them to direct their efforts towards the matters at issue. I have seen many cases in which psychologists had limited or minimal acquaintance with key issues in a case and overlooked them in their assessments or reports. It is difficult to conduct an optimal or proper examination when one does not know the specific purposes to which it should be directed. For example, exhaustive cognitive testing may be of little use if there is really no issue in the case related to intellectual functioning, or occupational functioning may not be covered thoroughly although it is far and away the most important area of dispute.

Monetary awards are intended to cover past, present, and future expenses, and to compensate for losses that follow and are due to the injury. Expenses might include damaged property, needed equipment (e.g., a cervical collar), alterations of one's home setting (e.g., a wheelchair accessible ramp), and treatment. Treatment might be one of the largest, if not the largest, component of damages. For example, in a serious brain injury case, extended inpatient rehabilitation and outpatient services may cost hundreds of thousands of dollars.

Losses might include earnings, pain and suffering, and, at least theoretically separate from the latter and where allowed, "hedonic damages." Hedonic damages refer to lost opportunities to experience pleasurable or positive events (as opposed to experiencing negative events), such as the enjoyment of attending one's senior prom. Consortium claims are also common, in which the spouse claims that the injured partner is less able to fulfill marital roles (e.g., the injured person has decreased sexual interest or capacity). With lost earnings, one considers time missed from work to date and whether the person has returned, or likely will return, to gainful employment. If the

individual has resumed working, decreased likelihood of promotions or ability to compete for higher paying jobs can be considered. In the case of younger persons, projections might involve the type of jobs the individual may have held in comparison to post-injury work capacities. In principle, damages should flow from the injury. An individual who previously had a poor marital relationship should not recover for the marital problems that pre-dated the event or were caused by other factors.

The above discussion should make it clear that individuals are generally not supposed to be compensated for having a condition per se, but from the consequences, problems, suffering, or dysfunction associated with the condition. For example, if two individuals both have Post-Traumatic Stress Disorder (PTSD), but one is usually able to control her symptoms, needs limited treatment, and is functioning well at work, she is likely to be compensated much less than the individual who has broad impairment and cannot hold competitive employment or maintain intimate relationships. This emphasis on the consequences of injury is one reason it is so important for an evaluator to go beyond a label and the immediate office setting and try to appraise, if scientifically viable methods are available, an individual's life functioning (see further below). Failure to do so can result in a serious over- or underestimation of damages. For example, an individual with orbital brain damage may seem relatively unremarkable on cognitive testing and structured interview but may exhibit incapacitating problems in everyday behavior, whereas another person with markedly elevated scores on personality testing and many complaints may be functioning proficiently.

Although it may fall within the psychologist's purview to describe areas of positive and negative functioning and to formulate projections over time, certain technical aspects of damage appraisal, especially their conversion to monetary values, are usually better left to others (which does not necessarily or always mean that one endorses their methods). For example, life care planners may list the equipment a brain-injured individual needs and the cost of various items, or an economist may translate conclusions about occupational capacity to numbers. Psychologists should be very careful not to undertake tasks for which they are not well qualified or knowledgeable or for which underlying methodology is dubious. As will be discussed, effective cross-examiners have refined abilities to detect weak points and to start tearing at these very locations. The credibility of an otherwise exemplary evaluation may be quickly demolished when the lawyer starts

asking the psychologist about the grades he received in business classes or technical questions about calculating interest rates and cost of living increases. It often only takes one or a few instances in which an expert has been shown to express opinions in areas she knows little about before the jury assumes the expert does not know much about anything.

Also, although I will not repeat the same point over and over throughout the chapter and instead will mainly emphasize practical consequences, there are compelling ethical reasons to restrain practice in relation to level of expertise and the availability of adequate scientific methodology/backing. Parties have much to gain and lose in legal proceedings, and the hope and expectation is that the expert in a branch of science will assist juries in reaching sounder decisions than they would otherwise, something that demands some reasonable level of knowledge and scientific underpinning.

A psychologist need not go to law school to become an effective forensic examiner, but should try to understand potential differences in the clinical and legal arenas, including the basic rules or principles of procedure and just what the issues are that define the scope of the expert's involvement and task. For example, in the clinical arena: (a) we are concerned first and foremost with the client's best interests (within the bounds of ethics) and we hope the client perceives us (or comes to perceive us) as such; (b) we usually assume that clients are not motivated to deceive us for purposes of self-gain, potentially at cost to innocent parties, and we usually do not obtain collateral records specifically for the purpose of checking on their veracity; (c) we may not be particularly concerned about cause, especially if we have narrowed possibilities down to the point that the alternatives do not change treatment; and, as follows, (d) we tailor our assessment to clinical needs and treatment questions.

In contrast, in the legal arena: (a) we presumably try to render an objective opinion, even if it might conflict with the examinee's self-interests, and we (should) assume that the examinee realizes we are a potential adversary (and ought to inform him of such); (b) we recognize that examinees might be motivated to deceive us and that we need to take active steps to check on their credence; (c) we often must focus on cause and may attempt relatively fine or subtle distinctions that may not impact on treatment at all; and, as follows, (d) we need to address, where appropriate and possible, key legal questions, such as everyday functional capacities, as opposed to what might otherwise be our main concerns (e.g., subjective beliefs and perceptions).

As is also evident, in a clinical context we do not typically expect our ideas or conclusions to be subjected to cross-examination designed to counter our view or to damage our believability or character; we might feel free to engage in hypothesizing and conjecture when trying to achieve greater understanding of our patient; and we do not expect the examinee to suffer potentially devastating consequences even if his claim is worthy but we cannot defend our opinions in court. Legal evaluators try to keep in mind the eventual uses of, and challenges to, their work. They apply this awareness to guide their efforts from the start, one hopes in the direction of performing high quality work that addresses the legally relevant issues and respects the bounds of scientifically sound method and opinion.

Finally, with all this said, juries' decisions may overlook technicalities that, in theory, should be decisive or at least highly influential. For example, although in principle the plaintiff usually must prove his case (the standard often being "more likely than not" or "by a preponderance of the evidence"), other considerations may prevail. For example, in one case, jurors interviewed after the trial agreed that they thought the plaintiff probably was not brain damaged, but that in case he might be they wanted to give him enough money to pursue the treatment he might require. In other situations, one or the other side may be perceived in such a negative light that judgments that should not be swayed by such reactions may nevertheless be altered. For example, the jury may be overly generous in deciding fault in favor of the defendant because they found her honest and likeable and thought that the plaintiff was an undeserving, devious individual. An expert who fares poorly or who is perceived as dishonest can in fact do considerable damage to her own side on issues that might seem only remotely related because she may cast doubt on others, such as the lawyer who retained her.

4.19.2.3 Treating Professionals Called into Court

4.19.2.3.1 Maintaining the treatment alliance as primary

My comments distinguishing the clinical and legal contexts assume, in the latter instance, that the psychologist has been asked to perform a legal assessment and knows from the start that this is the task. Treating psychologists may be called into court and, having served until then in a clinical role, often cannot be expected to have approached the case in a manner, or to have taken the steps, expected of a legal examiner.

Further, the psychologist might find certain expectations for legal examinations (e.g., taking little or nothing the patient says for granted and performing careful checks on credibility) contrary to the role of treater. Psychologists who testify in such circumstances usually are best served by openly acknowledging any limits in their methods. Any decrement in credibility associated with such limits may be more than offset by the believability often accorded to treating professionals as opposed to experts hired by one or the other side, and by the jury's relative intolerance for a lawyer who is overly aggressive with a treater, especially one who is likeable and candidly discusses shortcomings. For example, if the lawyer criticizes the psychologist for failing to obtain extensive background information, the psychologist might comment that were he conducting a legal assessment he would have sought this information, and that it might have been useful, but when the patient presented to him with suicidal ideation he was not worried about how he would look in court but about preventing a tragedy. In contrast, if the psychologist becomes defensive and insists the lack of background information is irrelevant or that such records could not possibly change his opinion, his credibility is likely to be damaged and he might no longer be perceived as the type of doctor the jurors themselves would seek out for treatment.

Treating psychologists may want to try to avoid testifying altogether, at least in cases in which they are concerned about damaging the therapeutic relationship. The specter of court can also compromise treatment relationships from the beginning. For example, parents who are considering divorce may be more concerned with impressing the psychologist favorably, just in case a custody suit eventuates and they need a helpful opinion, than about solving their problems.

When initiating assessment or beginning therapy with possible courtroom involvement looming in the future, and especially when circumstances are such that it threatens to impede progress, it can be helpful to take an immediate position that elevates the therapeutic role as primary and minimizes, to the extent possible, the likelihood that the present clinician will impact on any subsequent legal proceedings. A psychologist cannot assure a patient she will not be called on to testify in court and should think twice before assuring the patient that, even if called, she will not comply (because this could lead to serious sanctions, such as jail time). However, the therapist can say, with all sincerity, that she considers the treatment role to be primary. She can indicate that if asked to serve as a potential witness, she will tell the

attorney (or either attorney) that she has not attempted to form an objective or detached opinion and is likely to be biased or strongly influenced by the role as treater, that is, that she does not have, nor has she tried to develop, the type of impartiality desired of a testifying expert. She would therefore advise the attorney to retain an independent expert to perform an evaluation and address the issues relevant to the courtroom. Some potential clients, who, in fact, are primarily interested in using the clinician for courtroom leverage may abandon treatment immediately.

If the treating psychologist conveys such a therapeutically-oriented position to the attorney, he may well retain a separate expert. Attorneys have strong reservations, if not terror, about calling witnesses that are supposedly there to support their position but that they cannot control or predict. The damage caused by an adverse opinion is usually much greater if it comes from the attorney's own witness as opposed to that of the other side. The attorney may still call the therapist, not to state expert opinions, but simply to describe facts about such things as entries in the chart or treatment costs to date. Of course, it is difficult for the treating psychologist to arrive at a satisfactory solution sometimes, because a patient, although initially endorsing the psychologist's position of neutrality, ends up feeling disappointed or betrayed when the moment of truth arrives and abstraction becomes reality. Clients also need to realize that the therapist's commitment to treatment over legal involvement is not *carte blanche*, because certain behaviors or actions (e.g., child abuse) would not only likely alter obligations but are also associated with reporting requirements.

4.19.2.3.2 Release of assessment or treatment records

Although a series of generalizations can be provided about the release of records, only limited specifics are possible given both restrictions in the author's knowledge and the complexities and idiosyncrasies that often arise. For example, situations may occur regarding minors in state custody and who has the authority to release records; special questions may be raised when the mental competence of the patient is at issue, therapists may feel a strong obligation to review records in detail with former or present patients before releasing them; psychologists may be concerned about guarding the security of test materials, the clinician may be unsure what constitutes his file (e.g., does this also include records from other providers), or records may make references to others that are not party to a suit. Malpractice

cases raise an additional set of concerns; for additional details and guidelines, the reader is referred to Bennett, Bryant, VandenBos, and Greenwood's helpful book (1990).

Some psychologists believe they can assert a right of confidentiality when their records are subpoenaed in legal cases, especially if they or their patients never anticipated courtroom involvement, the legal issues seem indirectly or minimally related to the therapeutic work, and the file contains personally sensitive materials. On the contrary, confidentiality is a right of the patient, not the therapist. If the individual wants his records released (and especially if he is aware of the contents), and if there is no reason to believe the person's decision-making powers are seriously impaired, the therapist is obligated to comply. (Issues related to collateral records that the therapist has not generated or copyrighted materials, such as test items, can become complex and will not be taken up here.) Further, in civil cases in which an individual has placed his mental status at issue, the confidentiality of past or current therapeutic materials is likely to be waived, no matter what the client prefers (unless she drops her case or at least certain damage claims). There are exceptions, however. For example, some states have exceptions for alcohol or drug treatment, and juvenile records are often protected. Further, records may contain very sensitive materials that seem irrelevant to the case at hand but that could be terribly embarrassing to the client or might damage her public image and occupational endeavors. In such situations, the responsibility usually resides with the plaintiff's attorney to protest the release, or use, of the materials, although the keeper of the records (i.e., the therapist) may need to alert the appropriate party to the presence of sensitive materials in the file. A judge might choose to review the records privately (in camera), and then decide whether to release all, none, or parts of the material. As a general guide, the psychologist should not release records unless he has the consent of the patient or a court order (not just a subpoena) demanding their release, although one should not simply ignore a subpoena (see further below).

It is mistaken to assure a patient when beginning an assessment or therapy that she has an absolute right to confidentiality or that there are only a few circumstances in which records can be obtained by third parties. There are, in fact, many exceptions to confidentiality in most states, often over a dozen, and patients may end up disclosing materials that damage their courtroom cases or reputations that they otherwise would have withheld had they been properly informed. For example, an individual in a custody case may not want to tell the

therapist about occasional wild sexual fantasies that he can easily resist and would never pursue.

A few additional points can be noted. First, the psychologist should not just ignore a subpoena, and certainly not a court order. This does not mean that records must be released immediately, but if the psychologist is going to resist doing so or needs more information before acting, he needs to communicate the basis for the initial noncompliance to the relevant party in a timely and proper manner. In such cases it might be wise to consult an attorney. Second, it is usually wise to assume that treatment notes can be obtained in a legal case, or at least to recognize the real possibility that some other party might gain possession of the psychologist's records. Third, do not alter records. In a malpractice case, for example, it is not only obviously unethical and illegal to alter records, but it is often fatal to the case. Some psychologists seem naive about the mechanisms, and sometimes relative ease, with which altered records can be detected. Altered records demolish the clinician's credibility, and once this occurs the case is often, in effect, over. Fourth, if in doubt, contact an attorney, or at least another professional who is knowledgeable about legal matters. Issues relating to the release of records can quickly become complex, the adversarial nature of courtroom proceedings may be foreign to the professional, and the potential consequences of improper decisions can be serious, thereby calling for caution.

4.19.3 ADMISSIBILITY

The issue of admissibility, or what is allowed into evidence, has occupied many legal scholars and has many facets that often become extremely intricate. The issues that will involve us here is when, or whether, psychologists are allowed to testify as experts, and the extent to which their testimony might be constrained. To clarify the latter point, although a psychologist may be allowed to testify, the scope of testimony might be restricted. For example, in a criminal case, she might be allowed to address issues related to symptoms of PTSD that she believes the claimed victim of an assault manifests. However, she might not be allowed to address whether the examinee's presentation fits expectations for victims of violent rape, as such testimony might be seen as unreliable and as invading the province of the jury to decide matters of fact or ultimate issues. Various types of experts with various types of credentials testify in court. Psychological testimony is usually considered to fall within a branch of science, and thus the applicable standards are those relating to the admission of scientific evidence.

Standards for admissibility often vary across federal and state courts, and from state to state, and judges may vary considerably in the way standards are applied. What one court and one judge lets in, another might bar. There is little doubt, however, that the Supreme Court's recent ruling in *Daubert v. Merrell Dow* (1993), which involved the admissibility of scientific evidence, is having considerable impact. Many states are guided, in large or substantial part, by these new Federal guidelines. Briefly, prior to *Daubert*, admissibility of scientific evidence in Federal court was decided by two sets of standards: *Frye* (1923), which emphasized such considerations as acceptance within the scientific community, and a separate set of Federal standards. In *Daubert*, the court rejected *Frye*, elevated the Federal standards, and further elaborated upon them. Although the exact impact and interpretation of the Court's ruling will be clarified gradually as more and more cases are decided, at least two relatively clear trends have emerged.

First, *Daubert* explicitly acknowledges the judge's role as gatekeeper in deciding whether or not to admit scientific evidence. This means, in essence, that judges will likely feel freer to exclude expert evidence on scientific (or purportedly scientific) matters with less concern about reversal.

Second, in a number of cases, *Daubert* has led to the exclusion of evidence that might otherwise have been admitted under *Frye*, particularly in instances in which scientific foundations are weak (see Bersoff, 1997). In deciding whether to admit evidence and in evaluating scientific status, *Daubert* directs the judge to consider such matters as demonstrations (or lack thereof) of scientific validity, whether findings have been published in peer-reviewed journals, and whether the method has a known error rate. None of these considerations are necessarily dispositive, and the court did not attempt to create an exhaustive list of criteria nor to place them in hierarchical order. Thus, *Daubert* is very clear in situations in which all scientific indicators are either positive or negative, and increasingly incomplete in its guidance when indicators conflict (although the same basic point could be made for methods and approaches that practicing scientists and philosophers of science use to try to settle scientific disputes—see Faust & Meehl, 1992). As would be expected in such a circumstance, there has been, and undoubtedly will continue to be, considerable variation in the application and interpretation of criteria for deciding whether evidence meets the *Daubert* test. However, the weaker the scientific support, especially in extreme cases, the greater the

likelihood post-Daubert that evidence or testimony will be excluded. Consequently, in cases in which an expert, when challenged, cannot cite any decent scientific support for his opinion or assessment methods (e.g., there are no published studies), there is a very real chance that the psychologist will be prohibited from testifying. Overall, Daubert has reinvigorated challenges to the admission of scientific testimony, the potential testimony of mental health professionals and others is being placed under much greater scrutiny, there are an increasing number of decisions in which testimony has been excluded, and some reduction of junk science in the courtroom seems to be occurring.

What does all of this mean for the mental health professional? In an increasing number of situations, especially when the scientific foundation for testimony is weak, experts can expect challenges to the admission of evidence. In Federal court and in an increasing number of state courts, there is now more than a remote chance that testimony may be limited or excluded if there is minimal scientific support for the expert's opinions or methods, or if the literature is predominantly negative. Testimony might be allowed in areas in which scientific backing is stronger, but if foundations are weak across the board the psychologist may be barred from testifying entirely. Most commonly, an objection to the admission of testimony takes place before the trial. Written documents are submitted and, depending on the judge's discretion, there may be a pre-trial hearing at which arguments are heard and experts might testify.

In some areas there seems to be little question that psychological method or knowledge meets the Daubert standard. For example, many statistical methods have sound scientific foundations. There is also a good deal known about the potential impact of mild to severe brain injuries, at least in general (as opposed to the difficulties that might be involved in assessing specific impact in individual cases). In other areas, however, psychologists may be very susceptible to Daubert challenges. For example, many neuropsychologists construct their own idiosyncratic batteries, and there may be no peer-reviewed studies on the effectiveness of these batteries as a whole. Predictions of long-term outcomes may also lack scientific foundations, or studies of their accuracy may have produced consistently negative results.

Although a psychologist might view Daubert as an unenlightened or unjustified restriction, the hurdle that it creates—that evidence purported to be scientific should have a reasonable scientific basis—hardly seems outlandish. In fact, standards of this type would seem to be something that organized psychology, with its

stated commitment to science, could support. Such standards might ultimately improve psychology's position with the courts, given the methodological sophistication of many members of the field and the potential to build or extend strong scientific foundations in key areas of legal interest. For now, a psychologist might want to think carefully before agreeing to perform an assessment or provide testimony in areas with weak scientific underpinnings. Of course, this same argument certainly could be made whether or not there is likely to be a Daubert challenge. I would note in this context that although psychologists might believe that experience can partly or fully compensate for scientific shortcomings, even serious ones, the extensive literature on experience and accuracy raises major questions about this conclusion. In fact, the negative results of many studies on experience and accuracy in the mental health field have been noted by diverse individuals (e.g., Faust, 1984; Garb, 1989; Brodsky, *in press*, the latter being a revision of his earlier views [1991]).

Even if the courts allow testimony on some topic, a psychologist may feel that the state of knowledge is insufficient to develop trustworthy opinions and thus may opt not to be involved. Just as the courts may err in excluding testimony, they may err in admitting it; and it is a dubious position for a professional in a branch of science to assert that so long as it's good enough for the courts it is good enough for me. Given psychologists' methodological knowledge, situations certainly can arise in which the professional believes that a judge has overestimated the scientific standing of a test, method, or assertion and that, instead, knowledge is very shaky or dubious. Most psychologists' courtroom involvement is voluntary, and there is usually no external authority that compels the psychologist to testify if she does not wish to do so, or that will invoke official sanctions for nonparticipation. In these situations, psychologists can apply internal or professional standards in deciding on their courtroom involvement. (For a more extensive discussion of dimensions a professional might consider when determining a method's readiness for the courtroom, see Faust, 1993).

The court does not decide admissibility solely on matters relating to science. In addition, testimony must be relevant to the issue at hand. For example, a psychologist might be a first-rate authority on test bias, but the topic might not arise in the case, or it might be so secondary or remote to any of the matters in dispute that the judge excludes the expert. Additionally, the expert must contribute something over and above what the jury knows or can determine on its own. For example, if the defendant's left foot

has six toes and webbed feet and a plaster cast matches the pattern exactly, the jury probably does not need a footprint expert to provide commentary on the resemblance.

Experts also need to have adequate credentials. Usually, before opinions are expressed about issues in the case, the lawyer has the psychologist recite her credentials and then offers (proffers) the witness as an expert in some area (e.g., clinical psychology or PTSD). The expert should have credentials to support the offer. Psychologists may not realize that the standards judges apply when deciding whether to qualify the expert are often relatively minimal (e.g., a Ph.D. degree and some education and training in the area) rather than lofty, or that jurors may have difficulty separating, or care little about, differences in credentials that might strike the professional as weighty. A few publications might seem about as good as 20 to a jury, and a Ph.D. is likely to be an impressive degree whether or not it came from a big-name university. Ironically, some experts create major problems for themselves by failing to represent their credentials with a high degree of accuracy, when the difference often would matter little to a juror or to a judge evaluating qualifications. For example, an expert might exaggerate the number of patients seen with a particular problem, or might represent himself as a graduate of an APA-approved training program when, in fact, the program was accredited after the time of graduation and the listing is consequently inappropriate.

Take the last example. If the credential is listed accurately, the psychologist might be able to indicate, perhaps during his direct, that his program gained APA approval shortly after he graduated. The psychologist could further indicate that although the APA, in essence, approved the same program he attended, it would be technically incorrect for him to list himself a graduate of an APA-approved program, and he would not want to create a false impression. A juror is unlikely to give this a second thought, and it might even be a plus as it conveys honesty. In contrast, the following cross-examination illustration shows what can happen with even a seemingly small misrepresentation:

Question (Q): Doctor, is there anything you'd like to correct in your direct testimony before I start asking you other questions?

Answer (A): No.

Q: Doctor, is there anything about your credentials you may have stated in error?

A: No, not to my awareness.

Q: Doctor, do you remember raising your right hand before you started to testify?

A: Of course.

Q: And during that oath, you gave your word that you would tell the whole truth, isn't that correct?

A: Yes, I did.

Q: And doctor, you told us, did you not, that you were a graduate of an APA-approved program. Wasn't that your testimony?

A: Yes.

Q: That was a misrepresentation, wasn't it?

A: Absolutely not.

A: And your status as a graduate of an APA-approved program, that's the same representation contained in your resume, isn't it?

A: Yes.

Q: And this is the resume you send to lawyers, and have presented in court many times, isn't that true?

A: Well, I don't know about many times, but it is the information I have provided.

Q: It is the same information you would present to patients, would you not, if they asked you whether you came from an APA-approved training program?

A: I'm not sure a patient would ask such a question.

Q: Doctor, I'm sorry if my question was unclear to you. If asked by a patient, you would tell them that you were a graduate of an APA-approved program, correct?

A: Yes.

Q: Then let me ask you one more time. You are telling the ladies and gentlemen of the jury that it is true, without qualification, that you are a graduate of an APA-approved program. You are giving us your word, is that correct, just like you promised to be honest about the other areas in which you testified?

A: Yes.

Q: And, doctor, you graduated in 1985, isn't that the case?

A: Yes.

The lawyer can then produce an issue of the *American Psychologist* that includes the section indicating that a psychologist cannot describe his- or herself as a graduate of an APA-approved program unless the program was accredited at the time of graduation. Things can almost only go on a disastrous downhill course from there. For example:

Q: Having read that, and seeing the listing for your program and the date of approval as 1990, the truth is you are not a graduate of an APA-approved program, isn't that the case?

A: The listing is probably in error.

Q: Is that right doctor? I'll tell you what, we can check on the listings from other years of the *American Psychologist*, which I have on

the table here, and the sworn statement I've also obtained from the director of your graduate program. Do you think that will help clarify the truth of the matter?

A misrepresentation of this type can be fatal in not only the instant case, but in future cases. For example, in a subsequent case, a lawyer might ask the expert whether she has been busy calling all of the judges, lawyers, and patients with whom she's been involved to inform them that she misrepresented her credentials.

4.19.4 ASSESSMENT METHODS

In this section I will present various "tips" for performing legal assessments. Many of these guides apply similarly to conducting quality clinical assessments, although there are also a number of features distinct to forensic work.

4.19.4.1 Informed Consent

The fact that an evaluation is being performed for legal purposes typically does not alter the need to obtain informed consent before proceeding. The nature and purpose of the assessment should be presented, and it should be made clear that the clinician's primary obligation is not to advance the examinee's interests but to address certain questions, whether or not the conclusions help or hurt the individual's case. The expert should indicate who retained him, such as plaintiff or defense counsel, or judge. The expert should explain that his role is to perform an evaluation and not to provide treatment, and thus that there is no doctor-patient relationship. (This certainly does not mean the psychologist should not make treatment recommendations she deems appropriate or helpful.) At this point, the examinee may not consent to the procedure, which, in civil litigation, and often in criminal cases, is his perfect right.

There are cases in which experts are retained to perform evaluations, but with the understanding that they themselves may undertake treatment if they deem it appropriate. For example, the plaintiff's attorney may retain an expert she esteems to perform a neuropsychological evaluation of head trauma and, as needed, to provide rehabilitative services. The involvement in the dual role of forensic examiner and treator can create complications. For example, once engaged in a therapeutic relationship, it may be difficult, when it comes time to testify, to be as objective as one otherwise might be or to express views that will likely harm the plaintiff's case. Also, challenges can be raised about a financial

incentive to diagnose disorder in order to engage the individual in treatment for which one can charge. For these and other reasons, it is probably best, when possible, to avoid serving as both a forensic examiner and treating psychologist.

4.19.4.2 Clarifying Referral Questions and Deciding Whether they are Appropriate and Within the Clinician's Expertise

Lawyers usually make initial contact with an expert on a specific case by written correspondence or phone. Such communication will commonly eventuate in a request to perform an evaluation and perhaps some commentary about the questions the lawyer wants the expert to consider. At times, such requests are vague. For example, the lawyer might ask the expert to evaluate psychological status but not even specify whether she is primarily interested in cognitive or emotional status, or both, or neither.

Vague questions may reflect the lawyer's lack of familiarity with issues in the mental health field, and the expert should attempt to clarify just what is at issue in the case, such as the possibility of brain damage, emotional disturbance in response to a traumatic event, present and future work capacity, malingering of mental disorder, or whatever. The lawyer may be unsure about what a psychologist, or the expert in question, does and does not do and can and cannot do; and the psychologist may need to gain some familiarity with the file before a more productive conversation is possible or greater clarification can be achieved. For example, the lawyer may not realize that psychologists cannot prescribe medicine, that not all clinical psychologists are trained to evaluate neuropsychological status, or that many psychologists are not fully prepared to evaluate members of minority groups. Or, they might not realize that psychologists often provide treatment services, that many clinical psychologists are very familiar with various matters involving psychotherapy, or that the condition in question has a grave prognosis and that this issue would seem highly relevant to the plaintiff's case. It is not that the psychologist should do the attorney's job, but rather that the psychologist's input may be needed to hone in on relevant issues in the case or what is and is not possible to accomplish through a psychological evaluation.

Referral questions also need to be unobtrusive so that the clinician can determine whether the required scientific foundations exist to provide informed answers, and whether she has the requisite expertise. It is not only ethically

questionable to take on issues or questions for which one lacks the needed knowledge, but it can easily lead to courtroom fiascos. It is rather unnerving to have to admit on cross-examination in open court that this is only the second case of Fisbee's disease one has seen and that almost all of one's scientific knowledge on the topic was acquired through concentrated reading during the last week. This from the same expert who has contended, in effect, that the jury should defer to him over Dr. Jones, who has written the seminal work on the disorder and has acquired exhaustive knowledge about the condition through years of study. The psychologist is also likely to embarrass herself if a well-prepared and knowledgeable lawyer begins to ask about specifics. For example, someone who attempts to acquire quick knowledge about neuropsychology is unlikely to know what the abbreviations HEENT and GCS in the emergency room record mean, or may startle the jury when the response to the inquiry, "How many journal articles have you read devoted exclusively to mild head injury in the last 5 years," is, "None." There is a saying in aviation that there are old pilots, and there are bold pilots, but there are few old, bold pilots. This might be remembered as one engages in internal debate about whether or not to take a case. The psychologist might get by with questionable work for some time, but a single encounter with a skilled and well-prepared lawyer can inflict serious, long-term damage.

4.19.4.3 Access to Information

As will be seen, the quality of legal evaluations, and how well they stand up under scrutiny, often depend on the thoroughness of information gathering. Further, the time at which information is received can matter a great deal. For example, a psychologist who forms an opinion rapidly and in the face of inadequate information and who never revises his views in the slightest way, even after obtaining a great deal of additional information, may not seem credible. Also, a jury might look askance at an expert who reaches opinions and perhaps issues treatment recommendations, or even begins treatment, well before much of the information has been obtained, especially if that information was readily available. For example, many experts, who may have completed their reports a year or more earlier, indicate that they received or reviewed additional key information the morning of the deposition or the night before their courtroom appearance. It might be difficult for the jury to believe that the psychologist could be completely dispassionate

in her use and weighting of the additional information under such circumstances.

Further, given the uncertainties and ambiguities that often exist in our field, it would be surprising if considerable amounts of new information did not contain at least some evidence contrary to our conclusions, or even evidence that required new areas of inquiry or raised serious questions about the initial opinion. Receiving such information at a late date can therefore create major difficulties on the stand. For example, the expert might be asked a line of questions like the following:

Q: Don't you agree that mild head injury and the effects of serious alcohol abuse can look similar on cognitive testing? (If the expert disagrees, a statement he may have made in his deposition or in other cases confirming this assertion can be raised.)

Q: And didn't you describe the level of drinking that can result in abnormal performance on cognitive testing?

Q: And didn't you indicate that one of the ways you ruled out alcohol as a factor in Mr. Smith's testing results was his telling you that he had almost never touched a drop in his life?

Q: Until yesterday, when attorney Jones gave you the records, you were unaware of Mr. Smith's previous treatment for alcohol abuse, weren't you?

Q: So, when you formed your opinion and issued your report, you had not seen the treatment records describing Mr. Smith's history of alcohol abuse, had you?

Q: Surely this is the type of information you would have preferred to have known about when you conducted your evaluation, isn't that correct? (The lawyer is unlikely to care much, and may even welcome it, if the psychologist wants to fight her on this point, because it will probably seem unreasonable to the jury and further compromise the expert's standing.)

Q: And had you had this information, you would have asked additional questions about substance abuse, wouldn't you?

Q: You would have asked questions to find out whether Mr. Smith drank to the point that you would expect abnormalities on the testing, isn't that right?

Q: But it's too late to ask those questions now, isn't it?

An expert who denies the import of such questions may lose all credibility with the jury.

Many lawyers will not know what records a psychologist needs to perform a proper evaluation (see more on this below). Thus, it is up to the psychologist to inform the lawyer. The

mechanics for obtaining information (as opposed to a decision about what information to obtain) can be worked out in different ways, although it is often appropriate for the lawyer to do the needed leg work. It can be very difficult or impossible to obtain certain types of information (e.g., earlier school records in the case of an elderly individual), but it may become obvious that a lawyer is not really motivated to carry out the psychologist's requests and rather wishes to control the flow of information. This reluctance and attitude may severely compromise the quality of the evaluation and place the psychologist in a vulnerable position. For example, on deposition, an expert might be asked about the type of information she routinely procures, or prefers to have, when performing a comparable evaluation. Procedures followed in similar cases in the past might be raised. Discrepancies between the usual and present case can be pointed out, and the expert can then be asked to explain the reasons why the current evaluation is so much less complete, the extent to which gaps in information compromise the trustworthiness of the conclusions, and who has controlled the flow of information. A careless response to the effect that such extensive information is unnecessary can lead to such questions as, "Are you telling me that when you gathered far more extensive information in your many other cases and charged thousands of dollars for the totality of your time reviewing those materials, all that was unnecessary?" An expert who is true to the oath and admits that the lawyer exercised control might be asked whether lawyers taught their assessment courses in graduate school or whether it is a routine procedure when performing a psychological evaluation to consult a lawyer in order to determine what information the professional needs.

An expert might also contemplate whether she wants to be involved with a lawyer who tries to exercise this type of control. There is a big difference between a lawyer pointing out legal technicalities that fall within his purview (e.g., what standards of evidence will be applied or what exactly is being disputed in the case), and one who tries to control how the expert performs activities that fall within the psychologist's professional domain. One way to handle possible differences in viewpoint on such matters is to tell the lawyer that one will feel compelled to express reservations about one's conclusions given the present restrictions the lawyer suggests. Both parties might then make more informed decisions about course of action, including whether to part ways.

This does not mean that every difference, even a very small one, necessitates an all-or-nothing

stance, that the lawyer will never have a compelling counterargument, or that all such differences reflect a lawyer trying to control the expert. For example, a record a psychologist might ideally like to have may really be very remote to the issues in dispute and extremely difficult and expensive to obtain. Or, due to time limits imposed by the court, a psychologist may have to decide whether to perform an evaluation before all of the desired information is obtained. The psychologist might decide to go ahead, but to be sure the report contains an explicit statement describing the circumstances and indicating that opinions are subject to change or elaboration as more is learned. However, psychologists who take the long view probably should err on the side of caution. There are certainly some attorneys who wish to control access to information in order to increase the chances of a desired result, or who plainly want to manipulate the outcome. A psychologist will not only be on safer ethical ground but likely to have a much longer career as a forensic evaluator by avoiding such lawyers.

4.19.4.4 Design of Assessment Procedures: Some Do's and Don't Do's

4.19.4.4.1 Use the best available methods

It is important for psychologists to use the best available methods, not only to maximize the probability of reaching accurate conclusions, but because their work may be placed under intense scrutiny in legal proceedings. For example, the psychologist may be asked a series of pointed and specific questions about technical issues. Good attorneys can have a remarkable ability to pick up quasi-expertise rapidly, retain it, and generalize it to new contexts. An occasional attorney will ask questions about such matters as false-positive rates, standard error of measurement, or criterion-based validity with a proficiency that can shock an unprepared expert. Attorneys may also retain consultants to scrutinize records, educate them about technical matters, and provide them with background literature on the methods used in the case in order to better prepare for depositions and cross-examination.

Psychologists may use tests with poor normative information when better choices are available, or may use old or obsolete norms when more contemporary or complete normative information exists on the same test. For example, many neuropsychologists still use earlier norms for the Halstead-Reitan Battery (Reitan & Wolfson, 1993), which are decades old and which show a marked propensity to

overdiagnose certain groups of individuals, such as older and less educated persons. More recent norms of the type that Heaton, Matthews, and Grant (1991) developed, which adjust for age, education, and gender, can decrease overdiagnosis considerably.

Additionally, norms developed for one group of individuals may not be wholly applicable or may create major problems when applied to other groups of individuals. For example, norms for measures of motor speed developed with young individuals may be applied to the elderly, or norms developed among members of the dominant culture in the USA may be applied to recent immigrants with very limited English proficiency. Psychologists may be unaware of specialized normative data that have been developed for differing sociodemographic groups or, when such information is unavailable, may act as if the application of norms developed with one group to another group with distinct differences could not possibly raise any particular concerns or issues. When these types of problematic practices occur, lawyers can introduce materials from such sources as the Standards for Educational and Psychological Testing (APA, 1985), which describe the need for normative data appropriate to the individual under consideration.

For a substantial number of psychological tests, there are inconsistent sets of norms that may lead to very different interpretations. For example, depending on the norms used for the Halstead-Reitan Battery, the number of errors on the Category Test that falls two standard deviations beyond the mean may be less than 50 or more than 100 (see Faust et al., 1991). In the case of the Auditory Verbal Learning Test (AVLT) (Rey, 1964), a list learning task, many psychologists rely on the original norms developed many years ago in France with relatively small samples. There have been many subsequent studies showing that those norms are too demanding for many groups (e.g., Bolla-Wilson & Bleecker, 1986; Wiens, Crossen, & McMinn, 1988). A psychologist may be asked questions like the following on this topic:

Q: You told us that the plaintiff had problems on the Smith Memory Test, isn't that correct?

Q: And the Smith test is the one with the list of words, like a grocery list, correct?

Q: The standards you used to decide what was normal for the Smith Memory Test were developed more than 50 years ago, isn't that correct?

Q: Those standards were developed in another country where they speak a language other than English, isn't that correct?

Q: And there have been many studies published in the United States criticizing the use of those standards with today's Americans, isn't that true?

Q: There are more than 10 recent studies that would classify Mr. Smith's very same performance as perfectly normal, isn't that true?

At times, it is clear that one set of norms has been better developed than another, or it is obvious that the findings from one study deviate markedly from those of most or all other studies. Unfortunately, at other times, the underlying basis for conflicting norms, or the most appropriate choice among competing norms, is unclear, creating something of an impasse. A psychologist might consider the merits (or lack thereof) of using such tests, especially if tests with a better or clearer normative base are available that are designed to assess comparable areas.

Lawyers may also raise questions about other technical qualities of tests, such as reliability and validity. Some psychologists may see such questioning as meddlesome or something akin to badgering. However, standing on these technical qualities can be all important in assessing the likelihood of an accurate result, they are certainly fair game, and they often do need to be critically appraised. A psychologist may be surprised about the occasional lawyer's sophistication on these topics. For example, on deposition, one might be asked:

Q: In the language of your field, test-retest reliability refers to the stability of test scores over time, isn't that true? (If the expert is evasive on this and other questions, the attorney may call her own testing expert to provide appropriate and clear answers and to show, in effect, that the opposing expert was being less than genuine.)

Q: Isn't it a basic tenet in your field that tests that lack satisfactory reliability will also lack satisfactory validity or accuracy?

Q: Test-test reliability is often measured by the correlation coefficient, isn't that true?

Q: Correlation coefficients can range from .00 to plus or minus 1.00, isn't that correct?

Q: And in general, higher correlation coefficients are better, or indicate a higher level of reliability, isn't that correct?

Q: Isn't it true that (a prominent psychologist in the area of measurement will be named here) considers a test-retest reliability of .80 to represent a minimal standard? (If the expert does not acknowledge the author or source, there are other ways to get at the same thing, such as by asking what authorities on testing the expert respects, or about texts and

journal articles that were used in the expert's education and training.)

Q: You've not published your own text on psychological testing, have you?

At trial, after reintroducing these topics in clear and understandable language for the jury, the lawyer may proceed to show that various tests the psychologists used do not meet the standards for reliability the witness affirmed and thus, by the witness's own reckoning, cannot be trusted.

Of course, issues related to reliability are important not only because they can lead to courtroom embarrassment, but because adequate reliability is often necessary for reaching accurate conclusions. For example, deficient reliability can impede or cripple interpretive strategies that psychologists often consider essential, such as comparisons between test scores and pattern analysis. If reliability is low, the standard error of measurement of the difference $(SEM)_{diff}$ may be so great that there are likely to be frequent false-negative and false-positive judgments about whether true contrasts exist across tests. When comparing test scores, and especially when analyzing test score patterns, error is usually additive. Stated differently, the level of measurement error will exceed, sometimes by a great margin, the average level of error per test (with the margin expanding rapidly as the number of scores considered together as part of pattern analysis grows).

Test validity is not a global quality, and hence there is little meaning to a statement like, "The Fisbee Test is highly valid." Rather, validity is a specific quality, and a test that is highly valid for one purpose and with a certain population may lack validity for other purposes and with other populations. That is, validity refers to the interpretations given to test scores and not to the test itself. Therefore, when selecting tests, it is important to go beyond general information and to become familiar with research on specifics, in particular the specifics involved in the current application.

Particularized questions about validity are often essential to legal work because many of the issues psychological tests were originally designed to answer are different, in subtle or not so subtle ways, from those commonly asked in the courtroom. For example, in a criminal case, it may be of much greater interest to determine an individual's mental state at a previous point in time rather than in the present. The possibility of malingering and the need to evaluate for it are often considerably greater in the legal context. The differences that can exist in the purposes of traditional psychological tests and the questions that arise in the courtroom have led a number of

psychologists to develop instruments specifically designed to address legal issues, such as competency to stand trial (see Grisso, 1986) and malingering of psychological disorder (see Rogers, 1997a). Psychologists who conduct legal evaluations, especially in areas that traditional psychological tests are not designed to appraise and for which there is little or no research on these measures, might be well advised to become familiar with specialized methods, if they are not so already.

Along related lines, although one test may have usually outperformed another in comparative studies, important exceptions may exist in specific areas. For example, although Test A may be generally better than Test B, there may be much more extensive or positive research on the capacity of Test B to distinguish between the effects of alcohol abuse versus mild head injury. Such differences in level of background research might dictate a reversal of usual test selection. Such possibilities again suggest that specific knowledge about specific test qualities can greatly facilitate test selection.

Legal evaluations may also be weakened by the use of obsolete tests, particularly when more contemporary and improved versions are available. Some tests do not age well, and should the lawyer present some of the items in the cross-examination, jurors may be left shaking their heads and wondering, in fact, who these supposedly famous people that the test required the examinee to identify might have been.

Psychologists may substitute short, and clearly inferior, versions of full or standard forms of tests. Shortened versions commonly are far from perfect predictors of results on full versions of tests. As full versions of tests are usually rather imperfect predictors of the matters at issue, using shortened versions adds error to error (Ziskin, 1995). The lawyer might also be able to introduce statements by the test creator arguing vehemently against the use of short forms.

In some cases, alternate or parallel versions of a test are available, but one of the versions has been much more thoroughly researched than the other and there is inadequate research about the equivalence of the alternative forms. Especially if the test is administered only once, it would seem prudent to use the better known alternative. Similarly, for some tests, there are alternate administration formats, or alternate test materials or equipment, with one version being far better investigated than another. For example, a psychologist may use certain equipment when administering a test of finger tapping speed, but may depend on norms developed on other equipment when interpreting the results.

Although the use of one piece of equipment over another might seem like a trivial matter, research may show, as it does in this case, that different equipment can yield different results (e.g., Brandon, Chavez, & Bennett, 1986). The psychologist may also have to admit that he ignored the stern admonition of Dr. Reitan, a co-creator of the very battery he used, to avoid alterations in original equipment lest one compromise the interpretive value of results (Reitan & Wolfson, 1993).

I have consulted on more than a few cases in which psychologists used badly deteriorated test materials or made up their own "duplicates" of tests that contained errors in instructions or in the reproduction of stimulus materials. For example, in one case, the stimulus sheets for a task requiring color discrimination were more than a little faded; and in another, a homemade form for an intelligence test contained alterations in standard test instructions. Still another case involved Part B of the Trail Making Test (Reitan & Wolfson, 1993), a paper-and-pencil task, which contains numbers and letters randomly arranged on the page that are to be connected in order (e.g., 1-A-2-B-3-C, etc.). The stimulus materials are arranged so that the continuing line to be drawn from one number or letter to another never crosses over itself, thereby keeping the spatial layout relatively clean or simple. However, the psychologist had mispositioned the stimuli such that the respondent now had to crisscross over the line repeatedly, a new variation that seemingly made the task much harder. In other cases, reproductions of materials from personality questionnaires contained errors: it must have been difficult for the examinee to answer unintended alterations of original items from the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1951, 1983) such as the following: "I sometimes feel that there is a tight bank around my head."

4.19.4.4.2 Obtaining adequate information

One of lawyers' most common cross-examination tactics is to bring up concrete facts that contradict, or seem to contradict, the psychologist's opinion. For example, in a case of purported PTSD, in which the psychologist has described the plaintiff's decided tendency to avoid reminders of the accident, the lawyer may ask if the psychologist knows that the plaintiff replaced the totalled vehicle with a new one of the same make and model and regularly drives on the road where the accident occurred when an alternative route is easily accessible. In general, courtroom opinions will likely be much more vulnerable to cross-examination if the

psychologist lacks an adequate factual base. Reaching accurate conclusions about key matters also often demands such information. Consider again that legal evaluations often raise questions that differ from those that are most common in clinical settings and may well call for information gathering that differs in type and amount.

As discussed in Section 4.19.4.3, when information is received can be crucially important. For example, jurors are likely to react negatively if they find that the psychologist formed opinions very early and far before most information was reviewed, or waited until the eve of trial to perform a thorough analysis of records, years after issuing treatment recommendations intended to guide other professionals. The latter can look especially bad if the psychologist has insisted he places his role as a treator above that of courtroom expert and cares deeply about the examinee's welfare. The jury almost cannot help but wonder if this is impression management and disingenuous. After all, if the psychologist really is concerned about the individual, he would not have made important treatment recommendations on the basis of such incomplete information. Also, as noted, new records may call for entirely new lines of inquiry, and if materials are not examined until the 11th hour the opportunity may be lost. All of this again emphasizes the need for the psychologist to be thorough in information gathering in legal cases and, to the extent feasible, to obtain records earlier rather than later.

Information gathering usually must go beyond self-report and testing and include various types of collateral information, such as past and present medical records and reports, employment records, school records, and materials memorializing other's observations or reports about the plaintiff. It may be especially useful to access the observations of a neutral party who knew the plaintiff before and after the accident or event. These types of observations may be contained in background records (e.g., work evaluations) or in depositions, which may be as good or better than interviews. Information gathering will usually address at least four issues: a) prior functioning, b) current everyday functioning, c) possible causes or alternative explanations for the plaintiff's presentation or reported problems, and d) the accuracy and completeness of the plaintiff's report to the examiner.

To elaborate on these four areas, an analysis of prior functioning is important in order to determine the presence and possible extent of changes in functioning. In principle, plaintiffs are not to be compensated for problems that

pre-date the injury. Typical psychological assessment methods, by themselves, are often limited tools for determining prior abilities or adjustment. For example, psychological testing results usually provide, at best, inferential or indirect evidence. Much more direct information about prior functioning often is available or can be made available. Thus, extensive occupational records showing excellent and steady job performance are likely to be considerably more helpful and trustworthy in developing an understanding of past work function than inferences based on an IQ test. There are also, often, similarly direct and independent sources of information about current functioning that can supplement and extend, and serve as means to check on, impressions formulated from the plaintiff's report and testing.

Using collateral records to help appraise the accuracy and completeness of the plaintiff's report is essential, not only because there is a heightened risk of misrepresentation in legal cases, but because inadvertent misrepresentations can occur and lead to over- or underestimates of adverse changes. For example, depressed individuals may understate, or individuals with serious brain injuries may overstate, their current level of functioning. Also, the plaintiff may simply have forgotten, or may be unaware of, important matters. For example, she probably will be unable to state precisely when her developmental milestones were achieved, may be uncertain about or unaware of performances on past standardized testing, or may be confused about prior medical diagnoses. A psychologist is in an especially compromised position when he does not try to obtain collateral information in areas in which the deficits he assumes the examinee manifests would seem to impede that person's capacity to provide the very information the professional seeks. For example, the expert might ask a person, who he believes has serious problems in long-term memory, about remote historical events.

Many sources of information are usually potentially available to psychologists conducting legal evaluations. Past testing frequently can be obtained through such sources as school records, and sometimes through military records and pre-employment evaluations. Past work records, medical records, pharmacy records, psychological and counseling records, and criminal records may also provide important information or leads about previous strengths and weaknesses and about possible causes for presenting problems. One can examine these materials for the facts they provide, and also for direct and indirect indicia of prior functioning and capacities. Forms completed in the past, for example, may allow the psychologist to deter-

mine whether his impression about decreased mechanical writing quality is accurate. Information about post-incident and present functioning may be available from similar and other sources, such as work records and work samples. Further, some individuals who stop working return to educational activities or develop new hobbies. Contemporaneous school records may suggest that a brain-damaged patient over-reported current academic performance. Alternatively, they may show abilities to learn new skills and to engage in problem solving that raise very serious questions about the accuracy or meaning of low scores on measures designed to tap such functions.

The plaintiff's deposition may provide an extended sample of various cognitive functions, such as language use and the ability to attend to questions. I have been involved in cases in which plaintiffs with purportedly severe attentional and language comprehension problems answered hundreds of questions without a single one having to be repeated and seemed to have no difficulty understanding the lawyer's inquiries, despite the recurrent use of complex grammatical structure and relatively sophisticated vocabulary. Review of the plaintiff's deposition may also provide experts with important information that is not contained in other records, such as materials on past accidents or problems with the law.

Legal cases often involve questions relating to quality of life, day-to-day functioning, and work capacity. Psychologists may not habitually collect detailed information in areas that are directly relevant to these concerns. For example, an accident may have forced the injured individual to discontinue various enjoyable activities or hobbies. It is difficult to address impact on day-to-day functioning unless one develops a reasonably detailed description and chronology in this area. A courtroom evaluation in which the psychologist assigns a diagnosis, describes areas of maladjustment, often in abstract terms (e.g., Smith has elevated anxiety levels), and presents a general treatment plan may be of minimal help in understanding possible damages or the specific consequences of an injury. Too many legal evaluations are devoid of an adequate connection with the individual's day-to-day life and functions. It is usually much more helpful to know what the individual did before and can and cannot do now due to his reduced memory than to know the percentiles for his scores on memory testing.

In the area of work capacity, except in gross or obvious cases, it would often seem difficult to reach a well-grounded conclusion about possible return to former employment without a

reasonable understanding of prior job requirements. A general label or description may not be sufficient, and one needs to understand just what the job required. In one case, the plaintiff indicated he operated a machine that made paper bags, which the psychologist assumed was a rather simple matter. However, it turned out that the equipment was incredibly complex and often required intricate adjustments involving dozens of steps. Psychologists need not feel obligated to address work capacity, even if it is an issue in the case, should it involve matters with which they do not feel sufficiently informed or if they do not believe the required scientific backing is available to perform this analysis. (It would be much better to inform the attorney of this limit in advance, however, so that counsel can consider retaining an additional expert and is not unexpectedly stuck at trial with no one to address occupational issues.) The point is that if a psychologist is going to address work capacity, certain information would seem essential.

A well-rounded picture of an individual and an understanding of her functional capacities is also likely to touch on such topics as self-help skills, household activities and chores, interpersonal relationships, activities outside of the home and job, the state of the marriage, child care responsibilities and capacities, and travel activities. The examiner may also want to know who handles other responsibilities of major importance, even if they come up only occasionally, such as large purchases. There is something potentially inconsistent about a claim of gross reductions in mental ability and the fact that this same individual, with the spouse's full blessing, handled delicate and complex negotiations for the acquisition of the new home.

In this vein, the psychologist might be especially alert to, and actively look for, circumstances in which the benefits and disadvantages of being capable and incapable are reversed. For example, it might be to the plaintiff's advantage to appear quite disabled during the psychological evaluation, but quite capable when applying for a business loan. Consistency in presentation across such circumstances, even when the individual has a great deal to lose, suggests something about the genuineness of incapacity. In one case, a plaintiff in a psychological injury case was also embroiled in an unrelated custody battle. The plaintiff had undergone independent psychological evaluations in each case. On the evaluation conducted for the personal injury case, he endorsed many items in the pathological direction on the MMPI, but reversed these answers in virtually every instance in the

custody case: when it was to his advantage to claim disorder he claimed disorder, and when it was to his advantage to claim psychological health he claimed health.

4.19.4.4.3 Conduct technically proficient evaluations

It is not unusual for opposing counsel to obtain the psychologist's complete file, scrutinize it closely, and have her own expert review it as well. Technical problems with the examination may well be uncovered and cause difficulties.

Some of the more common problems I have observed in psychologists' courtroom work include errors in scoring psychological test items and in summing and transforming scores. I have reviewed cases in which psychologists made dozens of scoring errors or miscalculated standard scores by 30 or 40 points. In fact, the psychologist's errors may be close cousins to those he used to diagnose disorder in the plaintiff. The psychologist might then face a series of questions like the following:

Q: You told us that Ms. Smith made errors on simple math problems, and that this entered into your conclusion that she is brain damaged, correct?

Q: You made errors on simple math problems also, didn't you.

Q: You told us that Ms. Smith sometimes did not attend to important details when performing paperwork, and that this was another piece of evidence suggesting brain damage, isn't that true?

Q: You forgot to date a number of your test records, isn't that correct?

Q: Dating a test record might be considered an important detail when performing paperwork, isn't that correct?

Q: And you told us that Ms. Smith sometimes failed to follow fairly simple directions and that this was another indication of brain damage, correct?

Q: Doctor, the instruction you had to follow, "Stop testing after five consecutive errors," is not complex, is it?

Q: And you failed to follow this direction on multiple occasions, isn't that correct?

Q: Let me see if I have this right doctor. When Ms. Smith makes errors, they indicate brain damage. When you make errors, they're just errors? Strike that, I withdraw the question.

In other cases, psychologists may violate standard testing procedures without any clear or compelling rationale. A psychologist in one case

repeatedly terminated tests prematurely, calculated scores as if this had never occurred, failed to mention these alterations in procedure in her report, and had even published an article on the need to adhere exactly to standardized administration formats. Fortunately for her, the case settled before trial. There are certainly situations in which a test must be terminated prematurely or it is sensible to alter procedures (e.g., dividing a test that is preferably administered in one sitting into two sessions due to the onset of extreme fatigue part way through), but these departures should be recorded and described. What is hard to justify is altering standard procedures without any good reason. Psychology can be difficult enough without introducing unnecessary sources of error. Standard procedures may also be violated due to insufficient familiarity with prescribed methods. For example, in the cases I review, many experts seem unfamiliar with the precise rules for scoring the Visual Reproduction subtests from the Wechsler Memory Scale-Revised (Wechsler, 1987).

Failure to follow standardized testing procedures can make the psychologist look bad. Take, for example:

Q: Doctor, you told us that Ms. Smith demonstrated problems in delayed memory, isn't that true?

A: Yes, there were problems that seemed significant to me.

Q: And you administered the delayed version of the *Fisbee Memory Test*, correct?

A: Correct.

The lawyer might next ask some questions to help the jury get a clear picture of the test and the way delayed memory is examined. The lawyer then asks:

Q: And you told us that Ms. Smith scored in the borderline range on the test, didn't you?

A: Yes, that's a label I used, although her percentile was rather low, at about the 9th percentile, if I remember correctly.

Q: Borderline is the category between normal and abnormal, right?

A: I'm not sure I would say it that way, but I would agree with the basic thrust of your question.

Q: Had she gotten a few more points of credit on the test, the score would have been classified as low average, isn't that true?

A: That's true, but she didn't.

Q: And you have already agreed that according to school records, Ms. Smith was functioning at a low average level in a number of areas before the accident, isn't that true?

A: Yes, but I'm not sure that applies to her memory functioning.

Q: And one of the reasons you are not sure it applies to memory functioning is because we have no direct pre-accident memory tests, do we?

A: Well, I think some of the past tests reflect on memory abilities, but it is true I have seen no records of specific memory tests that pre-date the accident.

Q: Doctor, let's return to the *Fisbee Memory Test* that you gave to help determine the possible impact of the accident. Doesn't the test manual state that the delayed version is to be given about 30 minutes after completing the immediate version?

A: Yes.

Q: Doctor, do we usually remember things better over a longer or shorter period of time?

A: Shorter may be better, but not always.

Q: Doctor, you're not telling us, are you, that if I want to win a bet about when these jurors will remember the most about the trial, I should bet on next month rather than tomorrow?

A: I don't accept your analogy, and psychological phenomena do not always follow what might be considered common sense.

Q: So we're learning. In any case, when you administered the delayed version, you did not wait 30 minutes as the manual indicates, you waited 60 minutes instead, isn't that true?

A: Yes, but I don't think it makes any difference.

Q: Doctor, the instruction in the manual does not say it's fine to wait 60 minutes, does it?

A: No it doesn't, but many studies suggest that the amount of memory loss that occurs after 30 minutes and 60 minutes is similar.

Q: Doesn't the manual also state, and I quote, "The examiner should follow the instructions specified in this manual for test administration exactly in order to ensure proper comparison with the standardization group"?

A: I can't recall exactly.

Q: Well, I can show you the manual if you want.

A: No, that's OK, I'll take your word for it.

Q: Certainly it is possible, is it not, that a different result might have been obtained if delayed memory had been administered after 30 minutes versus waiting twice as long? (It probably will not matter what the psychologist says.)

A: It is possible, but I doubt it.

Q: And doctor, the reason to give the test as the manual specifies is so that we know how the individual performs when the test is given as designed, rather than having to guess how it would have come out if it had been given as the manual instructs, correct?

A: Those are your words.

Q: Doctor, I'm asking you. However, let's just move on. In a sense, doctor, you were measuring her on the 30 yard dash but made her run 60 yards, isn't that true?

A: No, I don't think you can make that comparison at all.

(The expert might think that but the jury is likely to believe that the analogy makes pretty good sense.)

Q: You waited twice as long as the prescribed time, and Ms. Smith obtained a borderline score, which you agreed was just a few points below the low average range, correct?

A: Yes.

I have also reviewed reports laden with factual inaccuracies relating to such matters as level of education, the date of the accident, the number of prior accidents, the number of children the plaintiff has, job history, etc. It may be difficult for a juror to believe that an expert can reach accurate conclusions about things that cannot be directly observed and that are complex—such as the area in which the brain is damaged or about the inner workings of the mind—if the expert cannot get simple facts straight, such as the day the accident occurred.

4.19.4.4.4 Give adequate consideration to alternatives

Research on clinical judgment suggests that diagnosticians can increase accuracy by waiting longer before reaching conclusions and considering alternative possibilities more actively (Faust, 1984; Faust & Willis, in press). In courtroom cases, the opposing lawyer will often present causal theories or explanations that contrast to those the expert proposed. An expert who has already made a systematic effort to evaluate and consider alternatives is more likely to be correct in the first place and better prepared to defend his position. For example, when the lawyer asks, "Isn't it true that excess caffeine intake can also cause symptoms of anxiety?" the response might be, "Although that is certainly true, the chart I constructed of caffeine intake and anxiety levels shows that anxiety levels often were high even when caffeine intake was low." The expert who has not carefully evaluated plausible alternatives might have to admit, instead, that she is not in a position to say whether, or the extent to which, caffeine might have contributed to the clinical presentation. The increased probability of an accurate conclusion also enhances the likelihood of providing true assistance to the trier of fact and may be critical in the examinee's care, whatever the courtroom implications. For

example, if the plaintiff's cognitive dysfunction is due to the onset of a schizophrenic disorder as opposed to mild brain damage, very different treatment is likely to be indicated.

(i) Alternative diagnoses, including malingering

Given the overlap in symptomatology across conditions, or the lack of specificity of many symptoms (e.g., anxiety, sleep problems, difficulties concentrating), clinical presentations often raise multiple alternative possibilities that require careful analysis and reanalysis of the positive and negative evidence. The defense expert who is leaning towards adjustment disorder may need to reconsider the chronicity of the condition and the substantial level of maladjustment. The plaintiff's expert who diagnoses PTSD may need to re-examine the absence of physiological arousal and the infrequency of intrusive thoughts. A neuropsychologist who is quick to diagnose a mild brain injury may have mistaken it for a depression that started earlier following disturbing life events. A psychologist who identifies characterological disorder as a basis for what he believes are false or exaggerated perceptions of sexual harassment may need to abandon this conclusion when collateral records for the 10 years prior to the reported events nearly all suggest good adjustment and interpersonal relations. Along these lines, various diagnoses are packed with assumptions about previous functioning, and experts may miss an excellent opportunity to check on their impressions by examining whether records about these periods conform to expectation. For example, someone diagnosed with hypochondriasis, which is usually assumed to be chronic, would be expected to have voiced multiple medical complaints in the past, and not just since the injury 6 months ago. If various past medical records show select, delineated, and realistic medical complaints (e.g., the finger did turn out to be broken when x rayed), it may be time to rethink the diagnosis.

In many legal cases, psychologists should be more complete in the assessment of malingering. For example, they may not collect any collateral information and may limit appraisal of malingering to interview impressions and a single test that the literature shows to have poor sensitivity. I have reviewed many evaluations in which the Rey 15-Item Test (Rey, 1964) was the only measure specifically used to appraise malingering. The sensitivity of this test is so poor that in cases of malingering, it apparently stands a worse chance of detection than a coin toss (see Rogers, 1997a). Other research suggests that it is difficult to detect malingering through typical interview methods and clinical

impression (see Faust & Ackley, 1998); and as Ziskin (1995) suggested, when a mental health professional asserts otherwise, the lawyer can ask, "Each time you've been fooled, you don't know it, do you?"

The psychologist who aspires to expertise in legal assessment should have a solid familiarity with the literature on malingering. Rogers' (1997a) edited book is a good starting point in this venture. The literature demonstrates the limits of traditional methods (e.g., interviews, many psychological tests), suggesting that: (a) individuals can manipulate results on a wide variety of tests, and in so doing may fool clinicians into overdiagnosing emotional or cognitive disorder; (b) it may be very difficult to detect face-to-face lies, especially by interview or impressionistic methods; (c) experience by itself does not ensure adequate detection capacities; and (d) the base rates for malingering may be higher than many psychologists believe (but lower than other psychologists believe) (see Faust & Ackley, 1998; Rogers, 1997a). For example, Reynolds (1998) asserts that "reasonable and thorough research indicates that at least 25% of cases of head injury in litigation involve malingering" (p. viii). Familiarity with the "negative" literature on clinicians' detection accuracy using more traditional methods is helpful because it directs us to take additional steps and because there has been such a recent growth in malingering detection methods—alternative or supplemental methods are available. This is a prime example of an area in which recognition of limitations in the field has had a very constructive impact by sparking intensive, productive research efforts.

The MMPI remains the most thoroughly researched test for malingering detection. MMPI indices and methods for malingering detection go beyond such traditional scales as F, L, and K, and are well described in various overviews of the topic (e.g., Greene, 1991, 1997). Differences in application with emotional distress versus brain damage claims are important to recognize, although the MMPI may be of considerable value in malingering detection with both types of cases.

There are also many specialized tests for malingering detection that are at varying stages of scientific development and of varying utility. One such example is forced-choice methods (Pankratz, 1988). Tasks are set up in which the examinee is instructed to select the correct answer from among two or more choices. For example, the examiner may read a string of digits to an examinee, then show her two written strings, and ask her to select the one that she just heard. A delay may be introduced before selections are made to increase the difficulty,

or perceived difficulty, of the task. With a dichotomous choice format, even an individual with no memory capacity should achieve about a 50% level of accuracy through random guessing. Some malingerers overplay the role, failing to realize that performance that falls substantially below chance requires knowledge of correct answers. Although such types of forced-choice methods may have high valid-positive rates (correctly identifying malingering when test results are positive), valid-negative rates (correctly excluding malingering when results are negative) are often poor (Rogers, 1997a). There are, however, rapid ongoing developments with this and other specialized malingering detection methods, creating good reason for optimism and making it important to regularly update one's knowledge of the literature.

Interview methods are also available, in particular the Structured Interview of Reported Symptoms (SIRS), developed by Rogers, Bagby, and Dickens (1992). The SIRS uses various detection strategies, the majority of which seem to capitalize on false stereotypes or general beliefs about mental disorder. A variety of studies from independent researchers on a range of topics suggests that the SIRS may have strong properties for malingering detection (see Rogers, 1997b).

Other recent studies have investigated whether knowledge of disorder and knowledge of the strategies that underlie the design of malingering detection methods help the examinee escape detection (see Faust & Ackley, 1998). This research is still at a relatively early stage of development, although it suggests the following tentative generalizations. With testing methods that use something more than the very simple types of detection strategies (like the MMPI and unlike, say, forced-choice methods), knowledge of disorder seems to be of limited helpfulness. In contrast, knowledge of detection strategies increases—sometimes markedly—the likelihood of escaping detection. With typical or unstructured interviews, however, knowledge of disorder may well be useful, and perhaps the most helpful element in successful malingering. If these conjectures turn out to be correct, it seems likely that the chances of detecting better prepared or more sophisticated malingerers may be greatly enhanced by using a combination of specialized testing and specialized interview methods (such as the SIRS), although one must worry about inflating the false-positive error rate. Optimal use of these methods, especially if multiple data sources are obtained, will likely be achieved through the development and application of formally validated decision rules or actuarial strategies for

data combination (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1997; Meehl, 1954). The idea that more information is better may already seem obvious, but considerable research in clinical decision making across multiple areas shows that intuitions about these matter are frequently misplaced and that there are many circumstances in which greater selectivity in data use and combination would result in greater accuracy (see Dawes et al., 1989; Faust, 1984; Faust & Willis, in press).

(ii) Considering alternative causative factors

Assuming a condition (e.g., brain damage or PTSD) is present, the question of what caused it can completely determine the outcome of a legal case, and hence is often one of the forensic evaluator's central concerns. Although some experts depend mainly or solely on self-report to determine cause, even the most honest plaintiff may not know what created her difficulties, or can make inadvertent errors in associating symptoms with events and conditions. If patients always knew what caused their problems they would not need doctors or mental health experts to perform differential diagnosis. Overattention to temporal sequence or focus on a salient event may lead to misattributions, as might faulty diagnoses by other providers. Factors may be operating that the individual could not possibly have been aware of, such as an unrecognized exposure to a toxin at the work site that causes a gradual, delayed reaction. Suppose such exposure leads to insidious decreases in reaction time and coordination, which in turn lead to a car accident that causes a very mild head injury. It is easy to see how causal confusion can arise. In other instances, of course, plaintiffs purposely mislead. This may be very difficult to detect, because their conditions may be real but may have arisen from another cause that they will hide or disguise. The chances of uncovering the deception may depend largely on obtaining sufficient background information.

Assessment results may also stem, in part or in whole, from transient or extraneous factors (for discussion of multiple possible confounding variables in brain damage cases, see Faust, 1995). Individuals are sometimes tested under poor conditions that obfuscate conclusions about more stable impairments. In one case, a nationally known expert, who had been retained by the defense, examined the plaintiff in a room in an airport. He concluded that problems seen on testing probably were not caused so much by the head injury in question but rather by exposure to jet fuel fumes during the examination. In another case, an individual set out very

early in the morning to drive the 200 miles to the psychologist's office, and ended up so tired he could not stay awake continuously during the testing. I have even reviewed cases in which examinees told the psychologist that they literally had not slept at all the night before, and yet the testing proceeded. Children are sometimes tested when they are ill and cranky or suffering from ear infections. In other situations, psychologists continue testing for 8 or 9 or 10 hours, and then write a report indicating that one of the plaintiff's main difficulties is rapid fatigability (while maintaining they have obtained "valid" testing results or results that do not underestimate capacities).

Many psychologists have very busy schedules, and cancelling a case slotted for a half or full day can create real headaches, but sometimes there would seem to be no reasonable alternative. Otherwise, the psychologist may have to answer questions like, "Doctor, do you ever tell your patients, before an important examination, to try to stay up all night so that they can do their best?" In contrast, there may be circumstances in which these types of accompanying problems and maladies are stable features of an individual's condition. For example, it might be that the plaintiff's physical discomfort, versus the brain injury, was the greatest contributor to performance difficulties during the examination. However, if the accident caused painful orthopaedic problems that are unlikely to remit, the adverse effects on cognitive performance may well be characteristic and typical of the individual's day-to-day functioning.

The many factors that may cause extraneous or transient alterations in cognitive, emotional, or behavioral status include, to name a few, sleep deprivation, pain, caffeine abuse, alcohol and other forms of substance abuse, medication side effects, and a myriad of independent medical conditions. Stressors and problems separate from the accident may also impact greatly on emotional or cognitive functioning. In situations in which these types of factors may have altered results, it can be very helpful to re-examine the individual at a later time. For example, if there is a question about the relative contributions of mood disorder versus brain injury to performance on cognitive tests, retesting after the depression has been treated successfully may help to parse these factors.

4.19.4.5 Interpretive Strategies

A detailed discussion of interpretive strategies is well beyond the scope of this chapter, and only a few general guidelines will be suggested. Some

of these guides may seem obvious and might be best viewed as reminders; I hope they will not strike the reader as too rudimentary. Other guides stem from the extensive literature on decision making, which provides many useful ideas and procedures. The decision making literature often has not infused writings on clinical practice. This is unfortunate because many of these principles, while very helpful, are counterintuitive, even for professionals, and hence are unlikely to be recognized or realized without direct exposure. Introductions to the decision making literature can be found in a variety of sources (e.g., Arkes, 1981; Faust, 1986; Faust & Wedding, 1989; Faust & Willis, *in press*; Wedding & Faust, 1989).

The recognition that interpretation is unlikely to be better than the data upon which it rests highlights the need for careful collection of information and adherence to proper examination procedures. Prudent data gathering can go a long way towards increasing the chances of reaching sound and defensible opinions.

It is hardly newsworthy to suggest that not all interpretive procedures and available methods are equally sound. Not uncommonly, various procedures are applicable to the same data and lead to conflicting conclusions. For example, there may be five or more procedures for judging cooperation with the examination or malingering, with some pointing in different directions. Further, there may be no simple way to combine or integrate these results, because they may directly contradict one another. If one decision rule indicates that there is brain damage and the other that there is not, both cannot be correct.

Obviously, when all decision rules coincide, there is nothing to resolve, but when they conflict directly, one must select one over the other (or defer judgment). At times this may not be too difficult, because the great bulk of the evidence, especially that of the highest quality, points in a certain direction; but in other cases the results are more ambiguous. For example, what does one do if the single best decision procedure argues for Conclusion A, but two other methods of more modest accuracy argue for Conclusion B, and one is not sure how these latter two methods operate in combination?

In these conflictual situations, a psychologist is much better off, to illustrate a relative extreme (and assuming all other things are equal), if the decision rule given preference has stronger research support, has provided a clearer result, is better suited to the examinee, and, when pitted against the other decision rule(s), has been shown to be correct more often. As the table tilts in the other direction, the chances of being correct decrease and vulnerability to

informed and skilled cross-examination increases. These and other issues involved with the combination of data are addressed in such sources as Dawes et al. (1989); Meehl (1973); and Faust (1984) (see also Faust & Willis, *in press*). As these authors note, considerable research suggests that it is not necessarily best to try to combine all of the data or to analyze complex configurational relations. Formal decision procedures (e.g., actuarial methods) that have been developed through proper scientific methods almost always equal or exceed the overall accuracy attained through these types of attempts at subjective or clinical data integration.

Interpretive methods may fail to consider base rates (see Meehl & Rosen, 1955; Wedding & Faust, 1989), or the frequency of events. For example, the accuracy of diagnostic signs and methods varies in relation to the frequency of a condition, and adjustments in the application of such indicators and in estimates of likelihood are likely to be needed as base rates change. To illustrate, as the frequency of malingering decreases across settings of application, the same score on a simulation index signals a lower probability of malingering, and hence cutting scores may need to be raised to avoid an unacceptable false-positive error rate. Along related lines, characteristics that are common among the general population may be described as indicators of pathology. Some such "signs" are part of clinical lore, and their use may not have been modified in accord with research. For example, a psychologist may assert that five points of subtest scatter on the Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981) is indicative of disorder, although research shows that most normal individuals equal or exceed this level (see Matarazzo, Daniel, Prifitera, & Herman, 1988). Given mental health professionals' skewed exposure to abnormal populations, it may be difficult to determine whether features seen commonly among patients are indicative of pathology or rather just common among individuals in general and thus poor discriminators of disorder.

I have reviewed many reports in which pedestrian human failings were stated as self-evident signs of disorder. When supposed pathological features include sometimes misplacing one's keys, occasionally forgetting where the car is parked, and irritability with misbehaving children, a good cross-examiner can have a field day. It is often necessary to consult epidemiological research and associated literature to determine just how often features occur among normal and abnormal populations, and also whether they occur with differential frequency among different forms

of disorder. The latter type of knowledge can be essential to differential diagnosis.

Overdiagnosis, or the tendency to see pathology that is not present or to overestimate its severity, may result from various sources. Overestimation of prior functioning may lead to overestimation of changes in functioning. False conclusions about prior functioning may stem from such sources as faulty history that the plaintiff reported or faulty methodology. For example, some psychologists believe that the single highest, or few highest, test scores provide a good estimate of prior overall or general intellectual functioning (the so-called "best performance approach"). Given normal variation in test scores, this method is almost sure to lead to overestimates, and not uncommonly gross overestimates, of prior functioning, especially among unaffected individuals. Given the average seven point spread between the highest and lowest subtest scores on the Wechsler Intelligence Scales (Matarazzo et al., 1988), use of the single highest score with normal individuals will result in an average overestimate of 15 to 20 points in prior Full Scale IQ (FSIQ). The result will be an estimated loss of 15 to 20 FSIQ points in an individual with absolutely no adverse change in intellectual functioning. This might be the difference, for example, between a score near the 90th percentile and one at the 50th percentile. The absurdity of this situation is that, on average, an individual will have to have gained over 15 FSIQ points to be judged to have retained his prior capacities!

Another basis for overdiagnosis is the use of inappropriate norms that set overly demanding performance standards. This problem is especially common when norms are not adjusted for sociocultural and demographic features or test bias. For example, a test may show pronounced age and education effects and yet a clinician, perhaps unaware of these findings, might apply norms developed on young, highly educated individuals to much older and less educated individuals. Also, in at least some normative studies, individuals with virtually any significant risk factor are eliminated from the sample, resulting in supernormal groups, or groups whose performances well exceed the general population (this being one reason why average IQs are often so high in normative studies on various neuropsychological measures). The result can be a mistaken belief that unaffected individuals have suffered a loss attributable to the event in question (e.g., a head injury) because their test performances do not meet the inflated standard set by the "normative" group. This can also lead to the initiation of treatments which may carry risks for target problems that do not exist. Overdiagnosis can also stem from

overattention to select weaknesses, insufficient awareness of the overlap between normal and abnormal populations, and failure to recognize normal variations within or across individuals, the latter of which was previously illustrated by reference to subtest scatter on intellectual tests. Normal individuals are rarely unremarkable or well adjusted or well functioning in all respects, and the tendency to conclude that an individual is aberrant due to even minor shortcomings creates a hurdle that few of us would pass.

Other psychologists underestimate or fail to recognize pathology, with explanatory factors sometimes representing the flip side of the same coin that leads to overdiagnosis. For example, psychologists may be quick to assume that problems pre-dated the accident. Weaknesses on cognitive or neuropsychological tests may be attributed to pre-existing learning disabilities, even without checking school records. Emotional problems may be assumed to represent Axis II disorders, and hence characterized as life-long difficulties as opposed to consequences of a more recent accident. In a case in which I consulted involving an individual with a moderate to severe brain injury, a psychologist passed off various potential indications of frontal lobe disorder, such as impulsiveness and marked difficulties with interpersonal relations, as indicators of a borderline personality disorder (BPD). However, this individual also presented with various features unrelated to BPD but positively associated with brain injury, only demonstrated characteristics of BPD that overlapped with those of brain damage, and demonstrated virtually nothing pre-injury to suggest any type of personality disorder.

Some experts are very quick to assume that individuals are malingering and that this explains their symptom presentation. In a toxic exposure case, a psychologist casually dismissed seemingly strong neurological findings as feigned symptoms. Ironically, as it turned out, although the psychologist was likely right that the toxin did not cause a problem, she was otherwise quite wrong: further medical work up yielded a nearly definitive diagnosis of multiple sclerosis and left virtually no doubt that the plaintiff had serious neurologic disorder all along. Underestimations of prior functioning may also lead to missed pathology. Finally, in some cases, experts misidentify chronic or permanent symptoms for transient ones. For example, in brain damage cases, experts may be too ready to assume that all of the plaintiff's problems are due to depression when, for example, the original injury was serious, the individual did not appear to have a low mood when evaluated, depression has fluctuated but impairment has not, and at least some of the

observed problems (e.g., perseveration and aphasic errors) are much more strongly associated with brain damage than with depression.

4.19.4.6 Preparing Reports

Reports may or may not be introduced as evidence at a trial. When they are, every word becomes a possible target for cross-examination, and thus they should be prepared very carefully. Although I will leave detailed recommendations about report preparation to others, I do have a few suggestions. First, sloppy reports with factual inaccuracies can create a very bad impression, even if the errors are not substantive. In one case, due to a typographical error and poor notes, a psychologist could not say whether a plaintiff, earlier in life, had fallen out of a rocker or had been struck in the head with a rock. Reports should also strive to provide a balanced representation of the case and of positives and negatives. For example, the report of an expert that repeatedly glosses over negatives that seemingly are evident and should not be ignored may be difficult to defend. Alternatively, the report may contain nothing but negatives. For example, when describing responses on a questionnaire or a depression inventory, the examiner may list only the unhealthy responses. This can become especially problematic when, in fact, the majority of the responses are positive. This creates an easy target for the cross-examiner, to wit:

Q: Doctor, in presenting before us today, you strive, do you not, to provide fair and balanced testimony?

Q: In gaining a complete understanding of an individual, strengths can be just as important as weaknesses, isn't that true?

Q: Interventions or approaches to helping a person often build on someone's strengths, isn't that correct?

Q: When describing the results of the Method X Depression Inventory, you shared Mr. Smith's responses to three items, isn't that correct?

Q: Each of these items suggested possible problems, you would agree with that, wouldn't you?

Q: And each of us, doctor, is something of a mixture of positives and negatives, wouldn't you agree?

Q: There are 20 items on the Method X Depression Inventory, isn't that right?

Q: We haven't heard anything about the other 17 responses, have we?

Q: On item number 3, doesn't Mr. Smith indicate that he gets as much enjoyment out of things as he used to?

Q: You didn't mention that in your report or your testimony, did you?

The lawyer then reads five more positive answers, and the expert may make some comment such as the following:

A: You're only mentioning the positive responses.

Q: And it would be very wrong to only present one side of the picture, wouldn't it?

Some experts write exceedingly long reports that address many minor or irrelevant matters. Again, every word and every comment is potential fodder for cross-examination, and thus including matters that are really not important to the task at hand is unlikely to be of much benefit and could create a problem. For example, such text might include various factual inaccuracies that can make the psychologist look foolish.

Reports may also include many extreme statements. A good cross-examiner usually welcomes extreme statements or exaggerations, because they create an easier target. For example, in one case, an expert did not just say that one method was better than another, but that there was a general consensus among psychologists that it was not just superior, but far superior to another method. This extreme claim was easy to deflate when literature was introduced by the very author of the test describing its limitations. In another case, a psychologist testified that even a very mild brain injury affects every single aspect of an individual's functioning and existence. Although the case was resolved before trial, in part because this psychologist's assertions were so vulnerable, it would have been a simple matter to confront the witness with the many normal test performances (which were entirely consistent with pre-accident measures) and to ask her whether there was any chance they could have been exceptions to her claim. She almost certainly would have said no, leading to a rapid self-destruction.

The exact wording of reports may be important and lead to unanticipated legal consequences. For example, in one case, an extremely bright neurologist wrote a conclusory sentence that included the phrase, "but for the accident." His intended meaning was that the condition would have occurred whether or not the accident occurred. However, in the legalistic world of the arbitrator, the conventional interpretation of this phrase was to the contrary, that is, that in the absence of the accident, the condition likely would not have occurred. It is generally much safer to stay away from legal terminology unless it is necessary and one understands exactly what one is doing, and

rather to just say things plainly and clearly. The wording of a report should also be checked carefully to avoid unintended meanings or interpretations, although it may be possible to clear these up at deposition or trial without creating a problem.

4.19.5 LAWYERS' STRATEGIES AND TACTICS

When dealing with adverse expert testimony, the lawyer's most basic task is to undermine credibility. Some lawyers will attempt to do this with a scalpel and others with a blunt stone, but the aim is the same in either case, and it is to the witness's advantage not to forget it.

Although an attack may be directed at personal matters (e.g., bias, financial incentives), this is often merely part of doing business and reflects nothing personal. The same lawyer who attempted to paint the witness as biased or incompetent may shake hands on the way down the courtroom steps and tell the expert he thinks she did an excellent job. On the one hand, most lawyers do not act in a blatantly hostile or obnoxious manner because they believe their case is best served if the jurors like them. A scowling, hateful, and abusive manner towards a witness, especially one who has done little or nothing to provoke it and who acts in a perfectly civilized and seemingly impartial manner, can hurt the attorney much more than the expert. For an expert who might be a bit thin-skinned, it can be helpful to perform a little personal *Albert Ellis* and tell oneself that such personal attacks often stem from a position of weakness and do not provide true commentary on one's human worth. A lawyer who is really loaded with ammunition may well prefer to take on an almost pained and solemn expression when bringing to light terrible problems with the expert's work (so that the jury does not end up feeling sorry for the doctor). On the other hand, arrogance and unwillingness to admit obvious points, and other similar demeanor may leave the jury hoping that the expert gets what he deserves and may give the lawyer license to deal out punishment. Obvious lack of preparation can also quickly alienate a jury.

If the lawyer is particularly successful in weakening an expert, she may then introduce materials and questions that are not so much intended to damage the expert's credibility further, for that job has already been accomplished, but rather to put on her own side of the case. For example, rather than the attorney attacking the credibility of a plaintiff directly, which, if overdone or too aggressive, could inflame a jury, it is often safer for an attorney to

do so through an opposing expert. The lawyer might raise inaccuracies in the plaintiff's report to the expert that seem purposeful and self-serving. A weakened expert is often fairly helpless against adverse, or seemingly adverse, evidence; and it can have a much greater impact for the attorney to introduce affirmative elements of her case through an opposing, rather than her own, witness. One might consider the effect when the opposing expert cannot fend off attacks on the plaintiff's credibility, versus the dubious impressions that can arise when the plaintiff seems credible and the plaintiff's experts, who vouched for the plaintiff's honesty, were not really questioned on this score. Rather, the first person to raise serious concerns about truthfulness just happens to be the expert the attorney hired.

Cross-examination often does not follow the contour of the expert's direct examination. Although, in some jurisdictions, cross-examination is supposedly restricted to the content covered in direct, in practice it is usually a relative free-for-all with few topics off limits. Further, a good cross-examiner rarely wants or needs to return to the points the expert covered *tit-for-tat*. Rather, the attorney is selective, looking for weaknesses or vulnerabilities. Many lawyers would much prefer to win a cross-examination 3 to 0, rather than 10 to 2. This is one reason experts should aspire to avoid weak components in their assessment batteries or procedures.

The areas of cross may or may not touch on any of the points covered in direct. In one case, a plaintiff's neuropsychologist had already admitted on deposition that during his entire professional career he had never concluded with relative certainty that someone was not brain damaged, nor had he ever directly identified someone as *malingerer*. He further admitted that in the instant case, he had presumed that the plaintiff was brain damaged nearly from the start, based solely on the very limited (and seemingly far from definitive) information he received at the time the referral was arranged. It would not matter very much what that expert said during his direct, because the lawyer was going to come back to these points, drive them home, and then stop, leaving the expert in shambles. In fact, a lawyer often is able to prepare a cross-examination without thinking all that much about what the expert will say on direct. Rather, the lawyer may consider the gist of the expert's testimony, and then focus most of her attention on the underlying bases or evidence for conclusions and on matters minimally related to the specific content of the direct.

Points of attack, to be covered in order, may include credentials, bias, flaws in the conduct of

the examination, questionable conclusions, and weaknesses in scientific underpinnings. Again, almost all of these subjects come back to the expert's credibility. Many lawyers would rather stay away from scientific topics, but others are well prepared to enter this arena. Experts may be accustomed to getting by with cursory or questionable answers to inquiries about underlying scientific methods or potential scientific weaknesses, and may be shocked the first time they confront an attorney who starts asking very specific questions about one or another line of research and will not settle for incomplete or vague answers. Some experts have permanently damaged their credibility by placing on record a host of patently wrong answers to questions about science. The result is the appearance of incompetence or, even worse, dishonesty. I have read more than a few transcripts with claims such as the following: Basic psychometric principles that apply to psychological tests are not relevant to neuropsychological tests; there is no type of board certification in psychology; the (you name it) battery is nearly 100% effective in diagnosing brain injury; there is an absence of research showing a limited relation between experience and accuracy, and so on.

4.19.5.1 Credentials

As discussed, most experts have credentials that sound impressive to a jury. For example, completing a doctoral dissertation, receiving an advanced degree, and publishing (at all) are likely to be viewed favorably. More advanced accomplishments are not likely to make much of a difference. It is ironic, then, that some experts will puff and inflate, or even distort, small points that a juror is likely to find unimportant or trivial. As illustrated earlier in the cross-examination exchange on graduation from an accredited program (see the heading, "Admissibility"), these little points can expand into disasters if they are discovered, especially if the expert will not yield in the face of obvious contrary evidence. It is not unusual for attorneys to conduct background checks and to obtain information about an expert's credentials and past experiences.

4.19.5.2 Bias

A cross-examiner may start by asking the expert whether he endorses the scientific method, and whether it calls for impartiality in examining data, including fair consideration of evidence for and against a proposition. It is difficult to imagine an expert answering negatively to inquiries of this type. The lawyer may

then try to show that the expert failed to live by such principles in the present case, or was not even-handed in considering and presenting evidence.

Some reports overemphasize or overattend to either the good or the bad, and may thereby make it easy for the lawyer to demonstrate bias. For example, an expert may list only the positive responses on an anxiety inventory. The lawyer might start reading the negative responses and after each one merely ask the expert, "Did I read the item correctly?" Alternatively, the expert may use dramatic terms and more lengthy narrative to describe weak cognitive performances, while downplaying, or even ignoring, strong performances. Problems may be referred to as "deficits," "impairments," or "serious shortcomings," but a remarkable performance is described as "essentially within normal limits."

Bias may also be suggested when experts fail to conduct sufficient investigation of alternative explanations for results. For example, an expert who quickly decided that the accident caused a brain injury may have been superficial or incomplete when looking into substance abuse as a possible alternative cause for reduced cognitive efficiency, despite records containing various suggestive references. Or, in a PTSD case, the expert may have made a perfunctory attempt to uncover and analyze other potential stressors. If the lawyer can show that alternative explanations were plausible but not pursued with anything approaching the zealotry displayed for the favored explanation, it can create a very negative impression.

Other experts come across as biased because they will not concede the obvious or will not give ground, even when they should. Similarly, some experts strike jurors as evasive. In one case in which I consulted, it was blatantly obvious that the plaintiff's signature was much poorer when produced at the psychologist's request than it was when he signed checks and documents in the course of his everyday life. When the lawyer merely asked, "Do these signatures appear to be of different quality?" the expert would not concede the point. This allowed the attorney to keep reframing the same basic question in a way that made the expert look increasingly outlandish, e.g.,

- Q: Isn't there a clear difference in the quality?
- Q: Isn't there some difference in the quality?
- Q: Do these different signatures look exactly the same to you?

The lawyer concluded the cross by stating, "Doctor, what would you say if I asked you ... ah, never mind, I know what you'd

say.” Although there may have been a perfectly reasonable explanation for the difference in writing quality (e.g., the plaintiff was heavily medicated when seen by the psychologist), the expert’s stubborn adherence to an unsupportable position convinced the jury that nothing he had said before, and nothing he might say after, was worth a second thought.

Other ways the lawyer might try to show bias include financial incentive (e.g., the expert charges much higher fees for legal than clinical work, or has performed numerous legal evaluations with the same attorney), or systematic error that consistently favors the expert’s position. For example, a defense expert may have repeatedly overcredited a plaintiff on cognitive tests, thereby underestimating loss in functioning.

4.19.5.3 Manner of Conducting the Examination

Examination procedures have been covered at length previously and will not be repeated here. Errors of omission and commission can create easy fodder for attorneys and can lead to a complete disregard of the expert’s testimony.

4.19.5.4 Erroneous or Questionable Conclusions

Erroneous conclusions can stem from various factors, such as mistakes in the scoring of tests or misapplication of scientific methods. In many instances, the correct answer is not cut and dry, and it may be difficult to show, conclusively, that an expert has made an error. Considering, however, the selectivity of cross-examiners, if even a few instances of clear-cut errors on non-trivial matters can be identified and brought out, it may greatly reduce the expert’s impact. Concrete facts often provide the lawyer with the best opportunity to find such occurrences, and the search may be greatly facilitated if the expert’s review of records has been incomplete or careless.

I have reviewed many cases in which, for example, an expert did not confirm educational level. For a neuropsychologist who adjusts most or all test scores in relation to education, this may lead to erroneous results on almost all measures. It can be difficult for a witness to regain his equilibrium with cross-examination surprises of this magnitude, especially if the normative tables are back at the office. I have also reviewed many cases in which the expert had taken a very limited occupational history. The psychologist might have testified that the plaintiff had never been fired from a job and had achieved a certain level of earnings, when work records show that these assumptions are plainly

wrong. Many lawyers are careful in collecting and reviewing facts, are attuned to incomplete or faulty factual renditions, and are capable of using such material to raise challenges, that is, they are much more at home with this type of subject matter than with psychological concepts and research. An attorney is likely to be considerably more comfortable and effective arguing about former earnings than about whether the subject-to-variable ratio was sufficient to conduct a multivariate analysis.

Experts sometimes present conclusions that sound just plain silly or out of touch. For example, they may make a great deal out of normal human failings or overlook obvious, everyday explanations for events. I have read many reports in which supposed deficits are illustrated through examples that apply to most anyone, e.g., difficulty getting organized for a vacation, occasionally forgetting the exact day of the month, or a tendency to fatigue by the late afternoon (this in a person with a hectic job and four children). Financial incentives in cases may be too readily dismissed, or not acknowledged as potentially relevant. For example, a plaintiff who complains bitterly about problems but who has complied with almost no treatment recommendations, even those involving minimal effort and possible discomfort, may be described as giving no thought to the legal case and only wanting to get better. Some psychologists are so used to thinking in complex and abstract ways that they tend to overlook more common or seemingly mundane considerations. At other times, experts may fail to think through the implications of presumed problems or deficits, and whether the kinds of things one consequently expects to be present in the plaintiff’s everyday life and lifestyle are present and those that seemingly should not be present are not present. For example, if an individual really has a severe problem with hostility and impulse control, it is doubtful his hunting buddies would continue their monthly get togethers at the cabin. Or, if the plaintiff really develops excruciating headaches when exposed to noise, one would not expect her to join a rock band. Everyday activities that fly in the face of the expert’s conclusions can be decisive with juries.

4.19.5.5 Scientific Status

An attack on scientific status might be broad and aimed at the field in general, or narrow and more specifically targeted at the particular methods used in the case. Many lawyers shy away from challenging scientific foundations, but others have little or no hesitancy to do so, are conversant with the issues, and have retained

a consultant to help them prepare. Further, unlike cross-examination, deposition inquiries about scientific status carry little risk, and helpful admissions or responses to even a small minority of questions may satisfy the attorney's eventual and basic purpose—to reduce or vitiate the expert's credibility at trial. An attorney may spend 2 or 3 hours asking many questions about the expert's methods and scientific backing, and then, at trial, focus on the one or few areas of questioning in this domain in which she feels she can make the most headway. Again, effective cross in just a few areas can inflict more general damage. Thus, other than wasted time and expense, it matters little if only one in three, or one in five, or one in ten areas of deposition questioning about research will be used at trial.

An expert makes the lawyer's search for damaging material on science much easier if she uses poor methods, lacks basic familiarity with pertinent research, makes grossly overblown claims, will not concede limitations that are well established in the literature, or repeatedly guesses when she is not sure about answers to questions. Much like certain forms of the martial arts, many points about science would probably carry little impact with the jury if experts did not take actions that gave them strength. Suppose, for example, there are a few negative studies on some otherwise well-supported method for appraising malingering. If asked, the expert could say, "Although most of the literature on the method is positive, one has to use it with some caution because isolated studies have not been supportive." However, on deposition, the expert might have argued that the literature is uniformly supportive. Once receiving that answer at the deposition, the lawyer may have nudged the expert out further and further on a limb. For example, the lawyer might ask about the importance of maintaining familiarity with literature on the methods one uses, how negative literature can bring a proposition into question, etc. Then, at trial, the lawyer can wave around the negative studies that the expert denied existed and recite the names of authors that the expert separately acknowledged as authorities.

As discussed, selection of the strongest possible methods is important not only in dealing with cross-examination, but in maximizing the chances of reaching correct conclusions and thereby fulfilling the presumed prescriptive (i.e., normative) role of a courtroom expert, to assist the jury in its deliberations. If there are no adequate methods available, or if the best available methods are questionable, the expert might decide not to undertake the assignment at all. Also, as noted, post-Daubert many courts are placing methods

under increased scientific scrutiny and eliminating those found wanting.

4.19.6 DEPOSITIONS AND TRIAL TESTIMONY

As part of the discovery process, most states allow attorneys to depose opposing experts in order to learn what opinions they may express at trial and the underlying bases for their views. Attorneys vary greatly in their approaches to depositions, and the type and position of the case may dictate strategy. Some attorneys play dumb, ask many open-ended questions, and try to get as much helpful material as they can while revealing as little as possible about their trial strategy and anticipated lines of cross-examination. Their aim is to surprise the expert at trial. Others are far more aggressive and ask pointed, challenging questions that are intended to inflict damage, even should the element of surprise be reduced. If the case is almost certain to go to trial, many attorneys will take a more guarded posture, trying to save their best material for the cross. If it is a case the attorney wishes to settle, the deposition style may be more aggressive and aimed at showing the other side that its expert has problems and that monetary demands or offers need to be adjusted. Many cases call for some type of strategy in between, for example, one that exposes weaknesses in a few areas to convey a message that might aid in settlement negotiations, but that reserves some or most of the material for cross should the case go to trial.

4.19.6.1 A Sampling of Deposition Topics

Although the supposed aims of a deposition may be to uncover trial opinions and their bases, the attorney often already has a very good idea what conclusions the expert will express, especially if a detailed report has been prepared. For instance, basic elements of PTSD cases are frequently similar, e.g., that it was the event in question that caused the disorder, that a decrease in level of functioning resulted, and that the plaintiff was essentially forthright and cooperative during the examination. The attorney's main deposition aims will probably lie in other areas, and the scope of questioning is often wide ranging.

Although many experts prefer to talk in more general and abstract terms (e.g., this is a serious case of PTSD that has caused substantial distress and diminished functioning), many attorneys prefer to talk in more specific and concrete terms. In the area of damages, they want to know specifics. What is it exactly that the plaintiff can and cannot do? How long will

therapy need to continue and at what frequency? Does the expert have an opinion about work capacity and about diminution in earnings? Exactly which problems can be attributed to the event and which pre-dated it? For problems that have supposedly been exacerbated, what is the precise extent of the change? Some psychologists become unnerved by these types of questions and react in ways that become problematical at trial. For example, they may fail to admit uncertainty, speculating about specifics that can be shown through concrete example and evidence to be wrong.

In addition to specific elements of damage, some of the topics that the expert can expect a well-prepared attorney to cover on deposition in a serious case can be discussed in turn. The attorney is likely to ask the expert about all her sources of information and when they were obtained. The attorney will probably request that the expert bring her complete file to the deposition, and may go through it document by document, asking when each was received. The attorney may also raise questions to determine who controlled the flow of information, e.g., did the expert request the document, did the attorney send it on his own, or did the plaintiff provide it? This type of questioning may translate into lines of cross-examination aimed at showing either that the expert formed opinions early and absent critical information; that the expert never has reviewed key documents; or that the lawyer, rather than the expert, determined what materials the expert reviewed. Further, knowing what documents the expert has not seen can give the lawyer a decided advantage. For example, if the attorney, but not the expert, knows that the plaintiff misrepresented her educational and occupational history, this can lead to embarrassing problems at trial.

As discussed already, the psychologist should not assume that the attorney who retained her will provide all relevant documents, or that it matters little when records were reviewed. Attorneys may not know which of the available documents an expert might want to see, or what new materials an expert might want to obtain. Some attorneys try to contain costs by limiting an expert's access to records or, in some cases, prefer to withhold certain documents for fear of their impact, hoping that the case will settle or that it will not create too big a problem at trial. Also, if a conclusion is reached weeks, months, or years earlier, and the expert does not review potentially critical documents until the eve of a deposition or trial, it can look very bad, especially if treatment recommendations had been issued. How can the expert explain why it took until 1997 to review documents that were

available in 1994 and that may have altered the diagnosis and treatment recommendations? Additionally, on deposition, the opposing attorney commonly asks the expert whether he plans to do anything further on the case before the trial, and requests that any changes in opinion be disclosed. (Some jurisdictions require amended reports if new or altered opinions are to be introduced as evidence at trial.) This puts the expert in a very difficult position if he plans a last minute review of documents he should have examined much earlier.

The attorney may also ask whether any literature was used or consulted in the case. If the response is positive, detailed questioning may ensue. Some attorneys will ask knowledgeable questions about such matters as norms, reliability, and validity. Related questions may be asked about the specific assessment methods used in the case. For example, for the various tests the psychologist administered, the lawyer may want to know what normative standards were used, whether other norms were available, the basis for selecting one set over another, what results would have been produced with other norms, and whether the norms contained demographic corrections. The expert might also be asked about the existence of literature that raises questions about her assessment methods or demonstrates limitations in their use, and whether each method is supported by a body of research on accuracy or validity. Again, a positive response may be met with numerous specific questions, such as what studies exist and who published them; whether they involved the exact same methods, populations, and questions; and whether there is contrary literature. With questioning such as this, inaccurate answers or overblown claims, rather than concessions about relative weaknesses, often cause experts far greater problems at the time of trial.

An attorney usually cannot introduce literature during cross-examination unless the opposing expert has acknowledged it as authoritative or as something he relied on in forming his opinion. (The attorney may still have the option of introducing that literature through his own expert.) Generally, the lawyer wants to find out at deposition whether the needed acknowledgement can be obtained, because it is highly preferable to know what is feasible in advance of trial and plan accordingly. Discovering at trial that an expert will not acknowledge an article that was to serve as the centerpiece of a cross-examination may leave the attorney on a bridge that just lost its undergirding. Once an expert acknowledges an article as authoritative at deposition, there is usually no keeping it out at

trial, even if the expert tries to backtrack. For example, should the expert say at trial, "Based on subsequent reading I no longer view that publication as authoritative," the lawyer can still refer back to the acknowledgement at the deposition and will probably be allowed to introduce the article and ask questions about it.

Some experts, due perhaps partly to uncertainty about the legal meaning of "authoritative" and because they wish to appear widely read, acknowledge a great range of literature as authoritative or provide very general endorsements. However, in the legal arena, authoritative means, in effect, that the expert defers to the source or considers it worthy of attention. This type of definition should be kept in mind when answering deposition questions about what is authoritative. Thus, for example, unless an expert believes that every article ever published in a particular journal is the definitive word on a topic or somewhere in this arena, it is probably a mistake to endorse that journal as a whole as authoritative. Rather, one might say something like, "The X journal contains a number of strong articles and others I do not think are as good, and I would really need to know what article or articles you are referring to in order to tell you whether I think they are authoritative or if I relied on them in this case." It is also reasonable to say that one respects a certain author, although one does not necessarily agree with everything the writer has said, and would need to know specifically what the attorney is referring to in order to answer a question about authoritativeness or possible use in the case.

Other experts will go to the other extreme, denying that anything is authoritative. Such experts can make a poor courtroom impression because they come across as pushing the view that there is only one person worth listening to on a topic, themselves naturally; jurors tend not to like individuals who act like self-appointed, self-anointed know-it-alls. Also, the lawyer often can still ask general questions about the literature that incorporate the gist of the findings, such as, "Isn't it true there are many studies showing (something contradictory to what the expert has described)? The lawyer might hold up a stack of articles to convey the impression that the assertion about the literature did not emerge from thin air. The expert may not acknowledge the authoritative status of the literature or the findings, and hence the lawyer might not be able to get further, or much further, into the specifics of the particular work. However, the lawyer has been able to bring the findings to the attention of the jury, and the expert's repeated rejections of contrary literature can create a strong suggestion of bias. The cross-examiner also has the option of putting on

her own expert, who can acknowledge the existence or status of literature the opposing expert has denied. Finally, it is almost always possible to get certain literature introduced through one or another means. For example, the lawyer will almost always be permitted to ask questions about the manuals for the tests that the expert used.

The attorney will probably ask whether the expert has talked with anyone about the case. If the answer is affirmative, questions will almost surely follow about the identify of the other party or parties and what specifically was discussed; and the attorney might decide to depose one or more of these individuals to get their descriptions. For example, if the defendant psychologist in a malpractice case talked to his supervisor, that supervisor may well be deposed.

The attorney may conduct thorough questioning about credentials. The expert may be asked about courses taken in graduate school, whether an APA-approved internship was completed, and if the expert pursued a post-doctoral fellowship. Questions may also be raised about supervisors and their qualifications, any malpractice claims or ethics complaints, performance on the licensing examination, continuing education activities, and board certification. There are also likely to be questions about research activities and publications. Some experts exclude publications from their resume that might be embarrassing, which tends to make things worse if the attorney uncovers them.

Attorneys frequently ask about fees and fee arrangements, and how charges for legal work compare to those for other activities. In some cases, experts insist that an evaluation was conducted primarily or solely for clinical purposes, but have billed at the much higher rate used for their legal work. The expert may also be asked how often she has been retained by the same attorney in the instant case, and by the attorney's firm, and perhaps what percentage of her income comes from these cases or her legal work in general.

The expert should be cautious about circumstances in which she is unlikely to be paid unless "her side" wins. For example, an expert may have a multi-thousand dollar fee outstanding with an impoverished plaintiff. A cross-examining attorney can bring out the situation and ask the expert whether these financial arrangements might make it difficult to be fully objective. The lawyer probably will care little about what the expert answers because the seed has been planted in the jury's mind, and an expert who denies such a possibility might well appear to be showing the very bias he denies he

manifests. It is thus often advantageous to have at least an understanding, if not a written agreement, that the attorney who retained the expert will be responsible for fees (although deposition time will usually be covered by the opposing attorney).

Experts may also be asked about possible alternative explanations for reported problems or difficulties, evidence that might exist for and against each possibility, and the process they followed in making their selections. They may be asked what type of evidence could be uncovered that might alter their selections or opinions. If the expert has not been thorough in reviewing background materials, the answer may contain elements that, in actuality, are present in the file. For example, suppose the expert answers that narcotic abuse could produce the symptoms he observed on his examination, and that the main way of distinguishing this possible cause from the one he identified is temporal sequence or association. It might just be that the plaintiff had obtained prescriptions for pain-killing narcotics from three separate doctors, was filling them simultaneously, and that the alterations the expert described started 3 months after the accident but just days after the visit to the last pharmacy.

The expert may be asked to critique the work and opinions of other experts, including those on her own side of the case. Sometimes the idea is to link a strong expert with a weak one, creating considerable difficulties for the former when the latter fares badly. If the expert says he endorses another expert's work completely, he had better be prepared to live with the approval not only in the present case, but perhaps in future cases as well. For example, failure to criticize weak methods when asked about views on the other expert's work may translate, in effect, into tacit endorsement, which can be brought up again in the next case and may also lead the expert into contradictory positions.

4.19.6.2 Some Suggestions for Depositions

Perhaps the most obvious suggestion, and one that applies similarly to depositions and cross-examination, is to prepare thoroughly. The required preparation varies and may include a review, or re-review, of the case records, relevant scientific publications, and, perhaps, test manuals and related materials. With large files, it would not be unusual to need a half day to refamiliarize oneself with the case, especially if it has been laid aside for some time, and for very large files, a full day might be required. With more complex files, constructing

a time line or chronology that summarizes major events or findings can greatly reduce the time that is needed to become reacquainted with the details at a later date.

It is usually sensible to meet with the attorney with whom one is working in advance of the deposition. If not already accomplished, the attorney needs to gain a clear understanding of what the expert can and cannot say, the boundaries of the expert's opinions, and what materials have been reviewed (e.g., the expert may have examined literature on his own that was not part of the attorney's case file). The expert also may need to be informed about particular technical or legal issues. For example, certain materials may have been excluded or deemed inadmissible and should not be referred to when discussing opinions. The expert may also benefit from learning something about the opposing attorney. For example, some attorneys may try to be especially provocative during depositions, hoping the expert will respond angrily and say something foolish.

If a lawyer is not interested in a pre-deposition meeting, it might be time to start wondering about the attorney and the situation that the expert might have gotten himself into. Similarly, the attorney needs to allow (and be willing to pay for) adequate preparation time. For counsel, this is likely to be one case with one expert, and the attorney's more basic obligation is to the client; for the expert, most every case becomes part of his "permanent record." The attorney is unlikely to lose the next case because her (now abandoned) expert in the previous case was trounced, but the consequences for the expert may carry across cases and years. Some experts, however, go overboard, spending too much time on minor or secondary issues or requiring inordinate amounts of time to prepare for depositions. It is hard to say exactly where to draw the line, and matters become much more difficult if disagreements about how to handle and prepare for a case arise late. If the expert feels that the attorney is setting restrictions that compromise professional standards, and if this situation is understood from the beginning, she can turn down the case as outlined. Alternatively, the attorney can be told what limits this could place on the soundness of opinions, and that any such shortcomings would be openly conveyed on a deposition. The attorney can then decide whether to loosen restrictions or retain someone else, the latter option sometimes being best for both parties.

Deposition questions, especially those that are well articulated, frequently call for specific answers and not dissertations. For example, if the attorney asks, "What is contained in your file?" or, "Did you consult any scientific

literature when working on this case?" an exegesis on the state of psychology is not required. It is often unwise to answer vague, unclear, or overly general questions. For example, the attorney may ask, "Do you believe Smith suffered a head injury?" It may be impossible to tell whether the question refers to any type of injury to the head (e.g., a facial laceration), or specifically to the brain. If one goes ahead and answers "yes," an attempt at trial to explain that one thought the question referred to any type of head injury, and not solely to a brain injury, may fall on deaf ears (assuming the expert ever gets a chance to attempt an explanation). Responding to overly general questions can cause similar problems. For example, the attorney may ask, "Isn't it true that individuals with PTSD show difficulties with interpersonal relationships?"

When problematical questions are raised, one can simply say that the question is difficult to answer as stated and explain why, e.g., "... because the answer differs depending on the specifics." If the question was not worded artfully but the expert thinks she knows what was intended, it is reasonable to respond with something like, "I understand your question to mean Assuming this, then ..." The lawyer, of course, can stop the expert if the rephrasing distorts the intended meaning (although he might prefer the question the expert has created over his original one).

Sometimes single or seemingly small changes in wording can ruin a question. For example, the lawyer might start by asking, "Is it fair to say that post-traumatic stress syndrome ...," and one might not know if the reference is to the formal diagnostic category or to symptoms that can follow trauma. Depending on which meaning is intended, the answer can change entirely. There is often nothing wrong with explaining what the ambiguity is and why it may be important to clarify the exact reference; and usually a good attorney can quickly discern whether experts are trying to be cooperative and are requesting needed clarification, or rather are being difficult and evasive.

Although it is important to listen to questions very carefully to be sure one knows just what is being asked and to determine whether the question is answerable, one can go too far and make the process miserable for everyone. It can be very difficult for lawyers who are not expert in an area to ask questions with a high level of technical proficiency and exactitude. As long as the question is clear or the expert, through rephrasing, can check to make sure the question is understood (with this assumed meaning memorialized on the record), it should be possible to answer. The relative ease with which

a deposition proceeds can depend in large part on the rapport between the opposing lawyer and the expert. If the lawyer is constantly rephrasing answers, attempting to get the expert to endorse a distortion that supports some contrary argument, or if the expert is really being nonresponsive in order to evade the import of questions, the whole process can bog down and become cumbersome, to say the least. If, however, the lawyer is doing his best to be clear and the expert her best to be responsive, and if the two can cooperate in clarifying ambiguities in questions, the process will usually go along reasonably.

The lawyer who has retained the expert will sometimes ask her to approach depositions parsimoniously, restricting herself to the question asked and volunteering no additional or extra information. As a somewhat hyperbolized example, if the lawyer asks, "Did you do anything to confirm allegation A?" the answer is, "Yes," not, "Yes, what I did was ..." Other lawyers will ask experts to answer more fully. They may fear that areas of testimony will be blocked if, upon questioning, opinions and their bases are not adequately explicated. For example, if cursory answers are provided to questions about the literature the expert relied on, she may be prohibited from discussing that literature at trial. After all, a fundamental purpose of a deposition is to learn the underlying bases for an expert's opinion, or to avoid courtroom surprises, so that the opposing lawyer can prepare for trial. Other lawyers want more complete answers because they hope or expect that a show of strength will give them settlement leverage.

Whatever the response style adopted, it is questionable to avoid an answer by exploiting trivial technical problems with the wording when the expert really knows what the opposing attorney is asking. I attended one trial in which an expert was asked, "Is there board status in neuropsychology?" The context of this question almost surely made it apparent to the expert that the lawyer meant board certification. However, because the lawyer did not ask the question exactly right and the expert wanted to avoid losing ground, the response was, "No." Aside from such a response style being disingenuous and arguably obstructing the legal process, getting caught can be costly. In the case in question, during a break, the lawyer recognized his mistake. He went back and made it very obvious to the jury that the expert had known exactly what the attorney was asking and had avoided answering on a technicality. This made the answer hurt 10 times more than it ever would have had it been surrendered earlier. Those less familiar with the trial process may be surprised

by the magnitude of the consequences that can follow when an expert is caught in an intentional misrepresentation, whether it is a direct lie or the offspring of evasion.

Uncertainty about the answer to a deposition question should be readily conceded. One cannot know everything, and such a concession is usually far less injurious than a wrong guess, or repeated wrong guesses. Lawyers may try to bait experts by insinuating in some way, perhaps simply by tone of voice or suggestive words, that they are asking about something that is very basic and that any professional would be expected to know. "Doctor, are you familiar with the large number of studies on . . .?" "Isn't one of the most frequently demonstrated findings in your field . . ." The lawyer may be hoping that the expert will make guesses, because with each guess there is an increased risk of error. If an expert guesses 10 times and is wrong three times, it is not hard to anticipate which three topics will be raised on cross-examination. The expert should not feel that she has to be absolutely certain about everything she utters in a deposition, but if the level of uncertainty passes some relatively low threshold, it might at least be noted (e.g., "If I remember right . . ."), or one should just say that one is not sure and does not want to hazard a guess. (Of course, if these types of uncertainties apply to opinions, and especially if the expert is merely guessing or almost tossing coins, he probably does not belong in the courtroom in the case; and one hopes the attorney who has retained him will not learn about this for the first time during the deposition.) Similarly, it is extremely risky (and potentially unethical) to make some claim that cannot be supported, e.g., "Despite the many negative studies on experience and accuracy, there are far more studies showing a positive relation between the two."

Depositions require a high level of concentration and can strain endurance. If one is too tired to pay close attention to questions, it is at least time to take a break, or to call a halt to the process. Stopping can be cumbersome if expensive travel arrangements are involved, but poor answers on depositions can destroy cases and haunt experts for years, or forever. Experts also should not let their guard down, something that is more likely to occur with fatigue. An off-handed comment on a break may be the first topic raised when the deposition is resumed. It may be particularly hard to maintain vigilance with a highly personable, friendly attorney. Whether or not the attorney is an honest, good-willed, considerate, and kind individual, her job is to win the case and she is likely to use anything she can (within professional ethics and the law) against the expert.

Thus, at least within the context of the litigation, that attorney may best be thought of as someone for whom you are the enemy.

Lawyers will sometimes spring new sources of information on experts during depositions, such as publications, medical reports, or documents relating to the plaintiff's everyday functioning, such as work records. If one is unfamiliar with the material, it calls for an open admission. If the lawyer wishes to ask questions, the expert should request the time needed to review the material. If it is not possible to perform an adequate review on the spot, the deponent should so indicate. Alternatively, if the expert believes she has acquired a reasonable grasp of the new material, she might state on the record that she will try to be helpful by answering questions about documents that she has just seen for the first time, although upon more careful study and reflection impressions or conclusions might change. Keep in mind that new materials might be presented out of context, which can lead to misimpressions. For example, one may come to learn that a letter describing insubordination as a basis for termination was written by a boss subsequently arrested for criminal behavior that endangered the public, and that it was the plaintiff who had heroically reported the misdeeds.

4.19.6.3 Trial Testimony

Prior to trial testimony, as with depositions, the expert typically should meet with the attorney. The meeting can address the topics to be covered on the direct examination and what might occur on cross-examination. Once again, the lawyer needs to understand what the expert can and cannot say and any reservations and uncertainties pertinent to the expert's views. Almost any attorney would want to know about these limitations and problems in advance, preferably as soon as possible, rather than discovering them during the expert's cross-examination. In one case, an expert testified that the plaintiff would have developed a particular disorder whether or not an accident had occurred. What he did not tell the attorney that retained him was that he believed it may have taken as long as 10 more years for the condition to develop if the traumatic accident in question had not occurred, something that came out at the end of the cross-examination. Had the lawyer known about this qualification in advance, he would have offered considerably more money to settle the case.

For each topic to be covered on direct, the lawyer needs to be able to ask the expert a question, or questions, that permit entry into the

subject matter. If a sufficiently precise question is not articulated, the expert may not even understand what the lawyer is asking, or may have no way to connect the question to the intended material. The lawyer can try out questions in advance to see if they prove sufficient to elicit the intended topic. An expert also needs to prepare for the possibility that upon objection, the judge may preclude queries or entire lines of questioning, which again emphasizes the need for expert and attorney to have a good idea about the topics to be covered. If the attorney does not really have this appreciation and is operating more by rote, unexpected alterations in the direct due to a judge's rulings or small slip-ups can be extremely disruptive.

Like a decent lecture, testimony is usually more effective if it is straightforward, accessible, unfettered with needless and endless complications, and not too lengthy. The jury may already have had a long day by the time the expert goes on, and, often, the expert's testimony is just one piece, although perhaps a crucial one, of a much larger composite that may include long strings of witnesses and days of evidence. A complex, obtuse, and exceedingly detailed presentation may quickly bore the jury and lead to inattention. This is not to argue for glossing over crucial points or simplifying at the cost of distortion. It is often very important to present the underlying bases for opinions, rather than just the conclusions, and in some instances detailed analysis and explanation are needed. For example, when discussing the use of a psychological instrument in malingering detection, a relatively detailed description of scales and their rationale may be necessary. Nevertheless, careful analysis will show that in most cases one's testimony revolves around only a few major points, many details are not of particular importance, and the direct can usually be limited to 1 to 2 hours or less. Also, every point raised in direct is potential material for cross-examination. Thus, unneeded complications or content are not neutral and primarily have a down side.

Visual displays can strengthen an expert's direct considerably by increasing interest and clarity. One might consider using at least one visual aid for each major point covered. It is usually preferable to keep demonstratives basic. A rule of thumb is to ask whether a visual aid is nearly, or completely, self-interpreting and can be comprehended quickly or with minimal explanation. Visual materials do not need to be fancy or elaborate, only clear. For example, a memory disorder in which rehearsal yields minimal gains in new learning can be easily illustrated through a learning curve. For visual

displays, I often use overheads because the equipment is readily available, materials are inexpensive, materials can be blown up to sufficient size with little distortion, and one can keep the lights up in the room, a real advantage late in the day.

Some experts try to impress juries with their technical knowledge. In particular, their presentations are laden with jargon. Although some technical terms may be necessary and have exact meanings that are otherwise difficult to capture succinctly (e.g., "post-traumatic amnesia"), other terms are pedantic and contribute little. Overuse of jargon is likely to alienate jurors.

On direct and cross-examination, the expert should remember which audience is the important one. It is not either counsel, but the trier of fact—the judge or jury—that will make the ultimate decisions and that the expert is there to address. If the opposing attorney hates you, or acts like you are an object of disdain, it does not influence the outcome of the trial an iota if the jury feels otherwise. Answers often should be directed at the jury, that is, one should look at the jurors and speak to them.

Many points about cross-examination have already been covered, and only a few additions will be provided here. Cross-examination, as noted, often does not follow the outline of the direct examination, so that the expert should not be surprised if completely different topics are raised. Depositions may give helpful clues about at least some of the upcoming points of attack, but there will almost always be some unanticipated questions at trial. As with depositions, or more so, preparation is extremely important. In addition to a pre-trial meeting with the attorney, one should be familiar, or very familiar, with the file, and with key background literature.

It is exceptional to get through an entire cross-examination untouched, and few cases that get to trial are so one-sided that at least some reasonable counterarguments to the expert's opinions cannot be raised. In the end, if the expert has done a decent job on direct, and if the attorney scores some points on the cross-examination but misses as or more often, it is the cross-examiner who has probably lost considerable ground. One of the worst characteristics of some witnesses is to not concede anything, even the obvious, as if the loss of even one exchange is intolerable and completely nullifying. An expert who is defensive and unreasonable, and who does not make what would seem to be required concessions loses credibility. If the attorney asks, in the context of *DSM-IV* criteria for malingering, "Would you agree that lying on the interview could be viewed as lack of cooperation

with the examination?" the simple answer would seem to be, "Yes." The attorney has not asked whether the expert thinks the plaintiff lied or whether the plaintiff was malingering. An expert might try to rush in to make these points, but a good attorney will simply say something like, "Doctor, I don't think you've answered my question. Please listen carefully," and then will slowly repeat the exact same question. If the expert again tries to dance around the question, she will begin to look evasive. When it seems to the jury that the expert would vehemently disagree should the cross-examiner assert that $1 + 1 = 2$, he becomes an object of ridicule.

The expert's responsibility is to conduct herself competently, professionally, and ethically, not to win or lose the case, and not to influence the outcome through sleight of hand. If honest concessions lead the jury to reject the expert's opinion, so long as these concessions do not stem from avoidable error or negligence (e.g., scoring errors), it does not mean that the expert failed to perform in a respectable and worthy manner. It may well be that the jury reached the right decision and that justice has been done. The issues at stake in the courtroom are often of great personal import, and one hopefully would much prefer to have one's opinion rejected than to prevail if, in truth, one was wrong. Almost every attorney wins and loses cases, and a "bad" outcome will not necessarily lead them to develop negative views towards an expert. This is unlike a situation in which an expert has been dishonest with the attorney, by withholding weaknesses in his opinions, for example, and gets caught on the stand. Unfortunately, in some such instances, the expert's actions may have destroyed the chances that the meritorious party will prevail.

4.19.7 REFERENCES

- American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, DC: Author.
- Arkes, H. R. (1981). Impediments to accurate clinical judgments and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology*, 49, 323–330.
- Bennett, B. E., Bryant, B. K., VandenBos, G. R., & Greenwood, A. (1990). *Professional liability and risk management*. Washington, DC: American Psychological Association Press.
- Bersoff, D. N. (1997). *The application of Daubert to forensic and social science evidence*. Presentation to the Federal Judicial Center's National Workshop for Magistrate Judges, Denver, CO.
- Bolla-Wilson, K., & Bleecker, M. L. (1986). Influence of verbal intelligence, sex, age, and education on the Rey Auditory Verbal Learning Test. *Developmental Neuropsychology* 2, 203–211.
- Brandon, A. D., Chavez, E. L., & Bennett, T. L. (1986). A comparative evaluation of two neuropsychological finger tapping instruments: Halstead-Reitan and Western Psychological Services. *International Journal of Clinical Neuropsychology*, 8, 64–65.
- Brodsky, S. L. (in press). *A hierarchical-conflict model of ethics in expert testimony*. Unpublished manuscript.
- Brodsky, S. L. (1991). *Testifying in court. Guidelines and maxims for the expert witness*. Washington, DC: American Psychological Association Press.
- Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993). 509 US 579, 113 S. Ct. 2786, 125 L ed 2d 469.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–1674.
- Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis, MN: University of Minnesota Press.
- Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology*, 17, 420–430.
- Faust, D. (1993). Use and then prove, or prove and then use? Some thoughts on the ethics of mental health professionals' courtroom involvement. *Ethics and Behavior*, 3, 359–380.
- Faust, D. (1995). Neuropsychological (brain damage) assessment. In J. Ziskin (Ed.), *Coping with psychiatric and psychological testimony* (5th ed., Vol. 2, pp. 916–1044). Los Angeles: Law and Psychology Press.
- Faust, D., & Ackley, M. A. (1998). Did you think it was going to be easy? Some methodological suggestions for the investigation and development of malingering detection techniques. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 1–54). New York: Plenum.
- Faust, D., & Meehl, P. E. (1992). Using scientific methods to resolve questions in the history and philosophy of science: Some illustrations. *Behavior Therapy*, 23, 195–211.
- Faust, D., & Willis, W. G. (in press). *Counterintuitive imperatives: A guide to improving clinical assessment and care by predicting more accurately*. Boston, Allyn & Bacon.
- Faust, D., Ziskin, J., & Hiers, J. B., Jr. (1991). *Brain damage cases: Coping with neuropsychological evidence* (Vols. 1 & 2). Los Angeles: Law and Psychology Press.
- Faust, D., Ziskin, J., Hiers, J. B., Jr., & Miller, W. J. (in press). Revision and update of Faust, D., Ziskin, J., & Hiers, J. B., Jr. (1991). *Brain damage cases: Coping with neuropsychological evidence* (Vols. 1 & 2). Los Angeles: Law and Psychology Press.
- Frye v. US (DC Cir. 1923). 293 Fed. 1013, 1014.
- Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, 105, 387–396.
- Greene, R. L. (1991). *The MMPI-2/MMPI: An interpretive manual*. Boston: Allyn & Bacon.
- Greene, R. L. (1997). Assessment of malingering and defensiveness by multiscale inventories. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 169–207). New York: Guilford.
- Grisso, T. (1986). *Evaluating competencies*. New York: Plenum.
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, 2, 293–323.
- Hathaway, S. R., & McKinley, J. C. (1951). *MMPI manual*. New York: Psychological Corporation.
- Hathaway, S. R., & McKinley, J. C. (1983). *Manual for administration and scoring of the MMPI*. Minneapolis, MN: National Computer Systems.
- Heaton, R. K., Matthews, C. G., & Grant, I. (1991). *Comprehensive norms for an expanded Halstead-Reitan Battery*. Odessa, FL: Psychological Assessment Resources.

- Matarazzo, J. D., Daniel, M. H., Prifitera, A., & Herman, D. O. (1988). Inter-subtest scatter in the WAIS-R standardization sample. *Journal of Clinical Psychology, 44*, 940–950.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.
- Meehl, P. E. (1973). *Psychodiagnosis. Selected papers*. Minneapolis, MN: University of Minnesota Press.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216.
- Pankratz, L. (1988). Malingering on intellectual and neuropsychological measures. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 169–192). New York: Guilford.
- Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery. Theory and clinical interpretation* (2nd. ed.). S. Tucson, AZ: Neuropsychology Press.
- Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.
- Reynolds, C. R. (1998). Preface to C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. vii–ix). New York: Plenum.
- Rogers, R. (Ed.) (1997a). *Clinical assessment of malingering and deception* (2nd. ed.). New York: Guilford.
- Rogers, R. (1997b). Structured interviews and deception. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 301–327). New York: Guilford.
- Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *SIRS. Structured Interview of Reported Symptoms. Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised*. New York: Psychological Corporation.
- Wechsler, D. (1987). *Manual for the Wechsler Memory Scale-Revised*. New York: Psychological Corporation.
- Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in clinical neuropsychology. *Archives of Clinical Neuropsychology, 4*, 233–265.
- Wiens, A. N., Crossen, J. R., & McMinn, M. R. (1988). Rey Auditory-Verbal Learning Test: Development of norms for healthy young adults. *Clinical Neuropsychologist, 2*, 67–87.
- Ziskin, J. (1995). *Coping with psychiatric and psychological testimony* (5th ed., Vols. 1–3). Los Angeles: Law and Psychology Press.