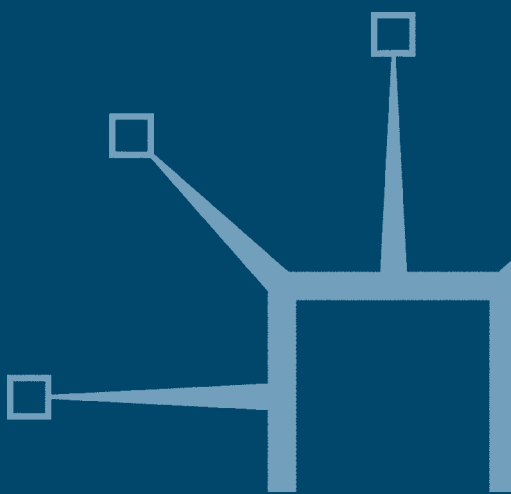
palgrave
macmillan

Psychology, Rationality and Economic Behaviour

Challenging Standard Assumptions

Edited by
Bina Agarwal and
Alessandro Vercelli



Psychology, Rationality and Economic Behaviour

This is IEA conference volume no. 142

This page intentionally left blank

Psychology, Rationality and Economic Behaviour

Challenging Standard Assumptions

Edited by

Bina Agarwal
Institute of Economic Growth, University of Delhi, India

and

Alessandro Vercelli
University of Siena, Italy

palgrave macmillan

in association with the
INTERNATIONAL ECONOMIC ASSOCIATION

First published in 2005 by
PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York, N.Y. 10010
Companies and representatives throughout the world.

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN-13: 978–1–4039–4253–1
ISBN-10: 1–4039–4253–6

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources.

A catalogue record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data
International Economic Association. Congress. (13th : 2002 : Lisbon, Portugal)
Psychology, rationality, and economic behaviour : challenging standard assumptions / edited by Bina Agarwal and Alessandro Vercelli.
p. cm.
Revised versions of papers presented at the Thirteenth World Congress of the International Economic Association, held in Lisbon in September 2002.
Includes bibliographical references.
ISBN 1–4039–4253–6 (cloth)
1. Economics – Psychological aspects – Congresses. 2. Choice (Psychology) – Congresses. 3. Rational choice theory – Congresses. 4. Altruism – Economic aspects – Congresses. 5. Decision making – Congresses. I. Title: Psychology, rationality, and economic behavior. II. Agarwal, Bina. III. Vercelli, Alessandro. IV. Title.
HB74.P8I58 2002
330′.01′9—dc22 2005043139

10 9 8 7 6 5 4 3 2 1
14 13 12 11 10 09 08 07 06 05

Printed and bound in Great Britain by
Antony Rowe Ltd, Chippenham and Eastbourne

Contents

The International Economic Association

A non-profit organization with purely scientific aims, the International Economic Association (IEA) was founded in 1950. It is a federation of some sixty national economic associations in all parts of the world. Its basic purpose is the development of economics as an intellectual discipline, recognizing a diversity of problems, systems and values in the world and taking note of methodological diversities.

The IEA has, since its creation, sought to fulfil that purpose by promoting mutual understanding among economists through the organization of scientific meetings and common research programmes, and by means of publications on problems of fundamental as well as of current importance. Deriving from its long concern to assure professional contacts between East and West and North and South, the IEA pays special attention to issues of economies in systemic transition and in the course of development. During over fifty years of existence, it has organized more than a hundred round-table conferences for specialists on topics ranging from fundamental theories to methods and tools of analysis and major problems of the present-day world. Participation in round tables is at the invitation of a specialist programme committee, but thirteen (now triennial) World Congresses have regularly attracted the participation of individual economists from all over the world.

The Association is governed by a Council, composed of representatives of all member associations, and by a fifteen-member Executive Committee which is elected by the Council. The Executive Committee (2002 – 05) at the time of the Lisbon Congress comprised:

President:	Professor János Kornai, Hungary
Vice-President:	Professor Bina Agarwal, India
Treasurer:	Professor Jacob Frenkel, Israel
Past President:	Professor Robert Solow, USA
President-elect:	Professor Guillermo Calvo, Argentina
Other members:	Professor Maria Augusztinovics, Hungary
	Professor Eliana Cardoso, World Bank
	Professor Duardo Engel, Chile
	Professor Heba Handoussa, Egypt
	Professor Michael Hoel, Norway
	Professor Jean-Jacques Laffont, France
	Professor Andreu Mas Colell, Spain
	Professor Kotaro Suzumura, Japan
	Professor Alessandro Vercelli, Italy

Advisers:	Professor Fiorella Kostoris Padoa Schioppa, Italy
	Professor Vitor Constâncio, Portugal
Secretary-General:	Professor Jean-Paul Fitoussi, France
General Editor:	Professor Michael Kaser, UK

Sir Austin Robinson was an active Adviser on the publication of IEA Conference proceedings from 1954 until his final short illness in 1993.

The Association has also been fortunate in having secured many outstanding economists to serve as President:

Gottfried Haberler (1950–53), Howard S. Ellis (1953–56), Erik Lindahl (1956–59), E.A.G. Robinson (1959–62), Ugo Papi (1962–65), Paul A. Samuelson (1965–68), Erik Lundberg (1968–71), Fritz Machlup (1971–74), Edmund Malinvaud (1974–77), Shigeto Tsuru (1977–80), Victor L. Urquidi (1980–83), Kenneth J. Arrow (1983–86), Amartya Sen (1986–89), Anthony B. Atkinson (1989–92), Michael Bruno (1992–95), Jacques Drèze (1995–99) and Robert M. Solow (1999–2002).

The activities of the Association are mainly funded from the subscriptions of members and grants from a number of organizations. Support from UNESCO since the Association was founded, and from its International Social Science Council, is gratefully acknowledged, particularly for specific help for the Lisbon Congress.

Acknowledgements

The Congress was held from 9 to 13 September 2002 in the Centro Cultural de Belém, Lisbon, at the invitation of the Ordem dos Economistas de Portugal, and was attended by 1,100 registered participants.

The Opening Session was addressed by the President of the Republic of Portugal, HE Senhor Jorge Sampaio, and by the newly-appointed Minister of Finance, HE Senhor Manuela Ferreira Leite; the IEA President, Professor Robert M. Solow, delivered a paper, 'Is Fiscal Policy Possible? Is it Desirable?' The programme comprised twenty invited lectures and three invited panels – on 'Growth in Developing and Transition Economies' (arranged by the Global Development Network), on 'Poverty Dynamics and Insurance' (organized by the European Development Research Network) and on 'The Turkish Financial Crisis' (prepared by the Turkish Economic Association). There were 198 contributed papers, a selection of which have been included with Invited Lectures in the four volumes of the Congress proceedings:

Bina Agarwal and Alessandro Verceilli (eds), *Psychology, Rationality and Economic Behaviour: Challenging Standard Assumptions*
Alan V. Deardorff (ed.), *The Past, Present and Future of the European Union*
Edward Graham (ed.), *Multinationals and Foreign Investment in Economic Development*
Robert M. Solow (ed.), *Structural Reforms and Macroeconomic Policy*.

Studies generated by the Global Development Network are published in Gary McMahon and Lyn Squire (eds), *Explaining Growth: A Global Research Project* (IEA Conference Volume no. 137).

The scientific responsibility for the selection of papers was in the hands of an International Programme Committee chaired by Robert Solow, with the following members:

Bina Agarwal, India
Maria Augusztinovics, Hungary
Victor Becker, Argentina
Miguel Beleza, Portugal
Enrique Bour, Argentina
Juan Camilo Cardenas, Colombia
Elinana Cardoso, Brazil
Vitor Constâncio, Portugal
Vittorio Corbo, Chile
Jacques Drèze, Belgium
Gene Grossman, USA
Seppo Honkapohja, Finland
Peter Howitt, Canada
Andrea Ichino, Italy
Firella Kostoris Padoa-Schioppa, Italy
Valery Makarov, Russian Federation
Andreu Mas-Colell, Spain
Mustapha Nabli, Tunisia
Ademola Oyejide, Nigeria
Adrian Pagan, Australia

Jean-Paul Fitoussi, France
Marc Flandreau, France
Augustin Fosu, Kenya
Jacob Frenkel, UK
Hans Gerbach, Germany
Luis Servén, USA
José Silva Lopes, Portugal
António Simões Lopes, Portugal
Hans-Werner Sinn, Germany
Kotaro Suzumura, Japan.

A National Scientific and Organizing Committee was convened by the Ordem dos Economistas de Portugal, under the chairmanship of its President, António Simões Lopes, who, with Amilcar Theias, Carlos Queiroz and Luisa Ahrens Teixeira (Executive Director of Mundiconvenius) formed an Executive Committee:

Mário Abreu
Luis Miguel Beleza
Daniel Bessa
Miguel Cadilhe
Teodora Cardoso
Eduardo Catroga
Maria José Constâncio
Vitor Constâncio
Vitor Pereira Dias
Erlânder Estrela
João Ferreira do Amaral
José Freire de Sousa
José Silveira Godinho
Manuela Ferreira Leite
Emâni Rodrigues Lopes
Isabel Almeida Lopes
Manuel de Oliveira Marques
Manuela Morgado
Isabel Almeida Mota
José de Almeida Serra
Francisco Soares.

The IEA is most grateful to the Ordem dos Economistas de Portugal, the Banco de Portugal, the Caixa Geral de Depósitos, the European Commission, the Fundação Calouste Gulbenkian, Portugal Telecom and other sources in Portugal which generously ensured most of the funding of the Congress. Among 46 other funders, mention must particularly be made (in alphabetical order) of the Banca d'Italia, the Bank for International Settlements, the European Central Bank, UNESCO and the World Bank. Cultural events were supported by the Fundação Calouste Gulbenkian for a concert at its Headquarters, by the Casino do Estoril for a Gala Dinner at the Casino, by the SECIL Corporation for a dinner for speakers at the Convento da Trindade. A Welcome Cocktail was offered on the opening evening at the Maritime Museum, Belém, and the publishers of the IEA conference volumes, Palgrave-Macmillan, gave a reception on the second evening to commemorate the Fiftieth Anniversary of the series, hosted by Amanda Watkins and Pooja Talwar. The Instituto Vinho do Porto provided a lecture and tasting of port wines. Day tours within the region of Lisbon were arranged in the three days after the Congress for participants and accompanying persons. The logistics of the Congress were efficiently handled by the staff of Multiconvenius, coordinated by Luisa Ahrens

Teixeira, its Executive Director. The staff of the Ordem dos Economistas do Portugal furnished additional assistance under the management of Carlos Quiroz.

The President of the IEA, Robert Solow, was Congress Editor. The IEA editorial team comprised Maureen Hadfield and Michael Kaser; the latter was responsible for the present volume.

List of Abbreviations and Acronyms

ABA	Advanced Business Administration Programme
AEH	adaptive expectations hypothesis
ARMA	autoregressive moving average
BBA	Basis Business Administration Programme
CEPR	Centre for Economic Policy Research (UK)
LR	likelihood ratio
MIT	Massachusetts Institute of Technology
MLRP	monotone likelihood ratio property
NBER	National Bureau of Economic Research (US)
NYU	New York University
PBE	perfect Bayesian equilibrium
RD	replicator dynamics
REH	rational expectations hypothesis
SDT	Shafir, Diamond and Tversky

List of Contributors

Bina Agarwal, Institute of Economic Growth, University of Delhi, India

Roland Bénabou, Woodrow Wilson School, Princeton University, USA

Friedel Bolle, Europa-Universität Viadrina, Frankfurt (Oder), Germany

Juan Camilo Cardenas, Universidad de los Andes, Bogota, Colombia

Ottorino Chillemi, Università di Padova, Italy

Florian Englmaier, Universität München, Germany

Ernst Fehr, Institut für Empirische Wirtschaftsforschung, Universität Zürich, Switzerland

Alexander Kritikos, Europa-Universität Viadrina, Frankfurt (Oder), Germany

Jochen Jungeilges, Universität Bielefeld, Germany

Oded Stark, ZEF, Universität Bonn, Germany

Theis Theisen, School of Management, Agder University College, Kristiansand, Norway

Jean Tirole, Institut d'Economie Industrielle, Université de Toulouse, France

Jean-Robert Tyran, Institute of Economics, University of Copenhagen, Denmark

Alexander A. Vasin, Moscow State University, Russian Federation

Alessandro Vercelli, Università di Siena, Italy

Yong Wang, City University of Hong Kong, China

You Qiang Wang, School of Public Policy and Management, Tsinghua University, Beijing, China

This page intentionally left blank

1
Introduction

Bina Agarwal
Institute of Economic Growth, University of Delhi, India
and
Alessandro Vercelli
University of Siena, Italy

Introduction

> 'How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it.'
>
> (Adam Smith, [1759] 1966: 3)

> '[T]he insistence on the pursuit of self-interest as an inescapable necessity for rationality subverts the "self" as a free, reasoning being, by overlooking the freedom to reason about what one should pursue.' (Amartya Sen, 2002: 46)

Economics today is at an exciting stage of evolution, as it begins to reopen routes of interchange with other disciplines. In this interdisciplinary exchange, psychology stands closer to centre-stage than most other disciplines. It has provided the grist for challenging many standard economic assumptions and catalysed the rapidly growing fields of behavioural economics and experimental economics.

For most part of the twentieth century, and especially since the 1950s, the characterization of the human being as *homo economicus* dominated economics. Strongly influenced by Newtonian Physics, to which is traced the formalization of modern economic theory, the underlying approach made the psychological characteristics of the economic agent largely irrelevant. Indeed, the approach acted as a protectionist barrier against insights both from other social sciences and the humanities, and from the observed complexity of human behaviour in real life. The last two decades, however, have brought an emerging recognition of the crucial role played

by the psychological attributes of economic agents in explaining economic behaviour.[1] This has paved the way for a less reductionist approach to the subject.

Under standard economic assumptions, human beings are presumed to be narrowly rational, motivated by individual self-interest, preoccupied with maximizing personal utility or satisfaction, driven by cold economic calculation without concern for others, capable of instantaneous learning, and so on. Real people are found to be more complex, driven not just by self-interest but also by altruism, guilt, liking, and other emotions;[2] caring not just about what they themselves get but also about what others get, and capable of forgoing personal payoffs for an equitable distribution of gains; wanting fair treatment and not simply higher incomes; influenced in their choice of workplace not just by considerations of profit or income maximization but also by emotions which tie them to a particular work environment and to fellow workers; learning slowly from mistakes rather than being able to self-correct instantaneously; affected by 'money illusion' which shapes their economic expectations;[3] capable of self-deception about their abilities and exuding levels of confidence (or lack thereof) inconsistent with their real talents; and so on. In other words, the introduction of psychology into economic behaviour provides a tool for challenging a central assumption of standard economics – that of *homo economicus* – in a way that strikes at the heart of microeconomic theory, where this assumption is embedded.

Indeed, as soon as we open the door of economics to psychology we introduce a breach in the traditional dyke that forestalls an analysis of the motivations and cognitive states of agents. (Or perhaps we should say, *reopen* the door to psychology, since in the nineteenth and early twentieth centuries a number of eminent economists – such as Nassau Senior, William Jevons, Irving Fisher, John Maynard Keynes, among others – did stress psychological factors in economic behaviour.[4]) Basically, psychology has pushed economists to recognize the heterogeneity of human responses and to examine a range of neglected factors that impinge on what motivates people and how individuals deal with complexity.

While many of these challenges can be explored theoretically, testing them empirically is equally important. Here too economists are now drawing on a tool which has long been standard in psychology – namely the use of controlled experiments to assess how people are likely to behave in given settings and contexts. Indeed, experimental economics has grown rapidly since the 1960s, and proved especially important in the development of behavioural economics. As is well-known, it is basically a tool to collect information by creating an artificial, controlled environment in a laboratory (typically a classroom) and using the tight experimental control to examine the relevance of psychological motives in economic decision-making. True, there are limitations to lab-based experiments, since the situations tested and responses obtained might deviate from the complexity of real-life situations

and responses in radical ways. But, arguably, this limitation can be overcome, at least in part, by setting up control experiments in the field among people facing similar real-life situations. This is now being done by some scholars (including Cardenas in this volume).

To these growing fields of behavioural economics and experimental economics, this book seeks to make a contribution. It brings together (in substantially revised form) a collection of papers that were presented as papers or as ideas at the Thirteenth World Congress of the International Economic Association, held in Lisbon in September 2002. All the chapters, in one way or other, focus on the insights that psychology provides in understanding economic behaviour, and the challenges it poses to standard economic assumptions.

The book examines these interfaces from several angles, but two ideas are especially dominant and constitute running threads in the volume. These are rationality and altruism. The interchange between economics and psychology inevitably raises, the question: What do we understand by rationality. The rationality entertained by mainstream economics is quite different from that entertained by psychology, and the tension between the two approaches was pointed out by some scholars a long while ago. Simon (1976), for instance, highlighted the divergence of vision and methodology between the two disciplines by characterizing the notion of rationality adopted in standard economics as 'substantive' rationality (emphasizing its unbounded character, its exclusive focus on optimizing equilibrium and self-interest maximization, and its narrow requirements of intertemporal coherence), and characterizing the notion of rationality adopted in standard psychology as 'procedural' (or 'bounded') rationality.

Another sharp critic of the narrow view of rationality found in mainstream economic theory – Amartya Sen (1977, 2002) – has emphasized the link between rationality and freedom. Rationality, he argues, depends on freedom not only because without some freedom of choice, the idea of rational choice becomes vacuous, but also because 'the concept of rationality must accommodate the diversity of reasons that may motivate choice' (2002: 5). Self-interest maximization can at best be seen as one among many goals that a person might choose to pursue, but a canonical selection of this one goal as an exclusive guide to rationality, and a rejection of all other motives and concerns that a person may have, would effectively involve, according to Sen (2002: 5, his emphasis) 'a basic denial of *freedom of thought*'.

Indeed, criticisms of narrow economic rationality have grown over the years, and the tension between the standard economics approach and the broader view of rationality emphasized in psychology, as well as in some heterodox contributions in economics, emerges in each of the chapters included here. Narrow rationality does not find corroboration in any of the analytical models and experiments contained in the book. Rather, several of the authors argue that apparently irrational behaviour can be seen as rational in the light

of the models suggested by them. Observed behaviour that appears to be inconsistent with a narrow view of rationality can still be considered rational from a broader and more comprehensive perspective. Extending the scope of what is deemed rational also helps extend the scope of economics itself, such as by relaxing the most demanding assumptions (mentioned above) that underlie the orthodox concept of rationality. In doing so, economics also reduces the gap between its own assumptions and methods and the assumptions and methods of psychology.

The second dominant idea that this volume explores is that of altruism. This can also be seen as related to the larger project of broadening what is deemed rational, beyond the single-minded pursuit of self-interest to the admission of other motivations, in particular 'other regarding' motivations. Several chapters emphasize the importance of altruism as one of the guiding forces of human behaviour – examining its origins, how it evolves, and what its implications are within the realm of the household, the workplace and the community. Altruism can impinge on many economic attitudes – 'the motivation to produce, the propensity to distribute, and the tendency to accumulate and transfer – within families, societies and across generations' (Stark and Y.Q. Wang, this volume, p. 85). It can also impinge on relationships between co-workers and between workers and employers, on incentive schemes, and on work contracts. And, in the context of communities, it can throw light on how institutions evolve or dissolve, whether or not people cooperate socially, and so on.

Discussions on these and related concepts are enriched by insights from other disciplines. In fact, the chapters in this book can be seen as a testament to the growing scientific exchange between economics and other disciplines and fields – the social sciences and humanities on the one hand, and the natural sciences (in particular, evolutionary biology) on the other. All the chapters draw important insights from psychology, but in addition some draw also on other fields such as cognitive sciences (Bénabou and Tirole), evolutionary biology (Stark, Y.Q. Wang and Y. Wang; Vasin), ethology (Vasin), epistemology (Fehr and Tyran; Vercelli), ethics (Jungeilges and Theisen), and political science (Cardenas). On the one hand, the chapters provide extensions, developments and new proofs of psychological ideas by using the tools of economics (game theory, principal-agent theory, intertemporal maximization, and so on). On the other hand, these contributions have the potential for influencing other disciplines, thus progressively broadening the scope of interdisciplinary exchange. Cases in point are the influence of microeconomics on evolutionary biology, apparent in the analogy being drawn between the selfish gene and *homo economicus* (Dawkins, 1990); and on neuroscience, apparent for instance in the argument that economic theory may provide an alternative to the classical Cartesian model of the brain and behaviour (Glimcher, 2003). In turn, an important theoretical insight in evolutionary biology, the 'Hamilton rule' (which relates to altruism between relatives),

is introduced in this volume by Stark and Y. Wang who clarify and extend its validity to economic issues by using a standard modelling device in economics, the prisoner's dilemma.[5]

All the chapters have sought to substitute narrow economic assumptions with more comprehensive ones. Some of the authors view this ongoing process of innovation in economic theory as a mere progression of standard theory; others see it as a paradigmatic shift. For instance, many chapters in this volume have modified standard utility functions by adding a term that takes account of inequality aversion, altruism and/or other social and ethical propensities. This modification could be interpreted as a generalization of traditional utilitarianism; or, as some authors contend, it could be seen as a paradigm shift away from the pure consequentialism and selfishness of traditional utilitarianism to a different vision, one that is open to social motivations and deontological principles. We leave our readers to judge for themselves which view they are inclined towards, after reading the chapters.

The book is divided into two parts. Part I explores the interface of economics and psychology theoretically, in terms of analytical models and methodological issues, while Part II uses the technique of experimental economics to explore this interface empirically.

Part I Analytical models and methodological issues

The volume opens with Bénabou and Tirole's chapter on 'Self-Confidence and Personal Motivation', which touches on many strands of the new literature that links economics and cognitive psychology. In particular the authors focus on the psychological trait of self-confidence and its effects on the behaviour and performance of economic agents. They argue that high self-confidence can have at least three types of values: a consumption value (a favourable view of oneself makes a person happier and so enhances her/his utility); a signalling value (self-confidence makes it easier to convince others that you have the qualities you believe you have); and a motivation value (it improves the individual's motivation and morale to persevere with her/his goals and overcome setbacks, thus improving performance). The authors note, however, that in many circumstances over-confidence can also damage performance.

These issues are clarified through a general economic model that seeks to explain why people value their self-image, and how they attempt to enhance or preserve it through various types of seemingly irrational behaviour, from self-handicapping to self-deception through selective memory or awareness management. As the authors show, self-deception in fact serves a number of rational functions. The suggested model of self-deception through endogenous memory reconciles to some extent the motivated ('hot') and rational ('cold') features of human cognition. By opening the door

to a wealth of problems and concepts typical of cognitive psychology, and adding new insights on both the psychological and economic features of human behaviour, the chapter extends the realm of economic analysis.

It does so through the judicious relaxation of a few crucial characteristics of *homo economicus* rationality. A case in point is intertemporal coherence – a crucial requirement of traditional ('substantive') economic rationality – which must be relaxed to study the game of strategic communication between an individual's temporal selves that may explain self-deception. This approach is inconsistent with traditional economic rationality but is consistent with a broader view of rationality that underlies the solution of the game. The insights so obtained could be applied to a wealth of unsettled questions in economics, such as the role of 'animal spirits' in investment, and (more or less optimistic) expectations formation in micro- and macroeconomics.

Alessandro Vercelli's chapter on 'Rationality, Learning and Complexity' also focuses on the impact of cognitive psychology on economic behaviour, providing a broad survey of emerging issues. Both casual observation and experimental research suggest that cognitive psychology significantly affects expectations and learning, which in turn play a crucial role in economic decisions. However, standard economics conceives of expectations and learning in a way that makes cognitive psychology irrelevant. The author clarifies the reasons for this neglect and seeks to specify the conditions under which the chasm between economics and cognitive psychology may be bridged. He argues that the crucial obstacle to closing the gap is the narrowness of substantive rationality, with its restrictive notions of expectations formation and learning. The chapter seeks to identify the main assumptions underlying the standard paradigm and to classify deviations from it in coherent alternative paradigms.

Vercelli makes a distinction between ontological complexity with respect to the properties of the economic system, and epistemic complexity with respect to the formal property of the model that represents it. He emphasizes the need to introduce more epistemic complexity, which he sees not as a virtue but as a necessity. In addition, he analyses the implications of different concepts of rationality in defining the theoretical and empirical scope of economic models. He claims that substantive rationality forbids any consideration of the subjective features of economic agents. Rather the concept applies only to an optimizing equilibrium characterized by stability, certainty or 'soft' uncertainty, and perfect reversibility of time in a closed and stationary 'world'. In particular the author shows that the standard assumption of rational expectations is inescapable in substantive-rationality models under uncertainty, but suffers from the same limitations. To take into account the complexity of the real economic world that is often characterized by disequilibrium dynamics, multiple equilibria, structural and dynamic instability and non-stationarity, it is necessary to introduce a more encompassing

notion of rationality. Such a notion would be one that enables, among other conceptual shifts, a more comprehensive hypothesis of expectations formation and takes into account the crucial role played by the cognitive and motivational features of economic agents. Only in this way is it possible to analyse the psychological determinants of economic behaviour, and make possible cross-fertilization between economics, psychology and the cognitive sciences.

Several of the chapters that follow, focus on a different psychological trait – altruism – and its importance in understanding economic behaviour. In their chapter, Stark, Y.Q. Wang, and Y. Wang explore both where altruism comes from and what its repercussions might be. This chapter is constituted of two papers that have emerged from the same research project, and which are presented here as Parts 1 and 2 of a single piece. Part 1 by Stark and Y.Q. Wang – 'On the Evolutionary Edge of Altruism' – focuses on how altruism evolves. The authors use the family, in particular siblings, as their starting point, arguing that the emergence of altruism within families can be seen as an important step for explaining the emergence and spread of altruism in society at large. They note that it is more likely for altruism to pervade large groupings if it evolves between siblings than if it fails to establish itself even within families. Accordingly, the authors discuss the evolutionary foundations of the emergence of altruism between siblings based on the 'Hamilton rule' suggested by evolutionary biology. According to this rule, altruism is likely to spread in a population if the benefit obtained from giving, times the coefficient of relationship, exceeds the cost of giving. This maximizes the replication opportunities of common genes, since the coefficient of relationship measures the probability of the genes being the same. Within a family, altruism would thus evolve if the benefit to one sibling from receiving help exceeds twice the cost of providing help borne by the other sibling, given that the coefficient of relationship between two siblings is one half.

Part 2 of this chapter by Stark and Y. Wang – 'The Intergenerational Overlap and Human Capital Formation' – already assumes that altruism exists (rather than explaining why it evolves), and the focus is on the economic consequences of altruism, a crucial repercussion being the level of human capital formation. The authors explain the strong positive correlation between the formation of human capital and life expectancy on the basis of parental altruism and the duration of the intergenerational overlap. Since education costs less if it is financed by parents than by market borrowing, the longer altruistic parents live the more will be the children's human capital investment. An extended overlap entails the formation of a larger quantity of human capital. This also explains the positive correlation between education and health. However, the authors note, this effect is separate from the returns to human capital – a higher life expectancy increases the period over which the returns can be reaped.

Although in this research Stark and Y. Wang primarily trace the repercussion of parental altruism on human capital formation, the altruistic trait can also have other implications. For instance, it can impinge on intergenerational transfers of income and resources, as well as on intergenerational transfers of the altruistic trait itself or of substitutes for altruism through the 'demonstration effect', namely the provision by adults of care and attention to their own parents, aimed at instilling appropriate preferences in their children (Stark, 1995). Of interest, too, is the possible extension of altruistic behaviour beyond the family to the larger society, in whether individuals cooperate or defect when interacting in non-familial groups; or in the formation of long-term time preferences and the determination of optimal consumption (Falk and Stark, 2000). These aspects have a central bearing on contemporary concerns such as environmental preservation (for example, do you use up most of a forest now or save most of it for your children?), and have promoted other fruitful interdisciplinary exchanges, such as between economics and political science.[6]

The next chapter, 'Human Reproduction and Utility Functions' by Vasin, like the work of Stark and Y.Q. Wang (Part 1 of Chapter 4), also draws upon evolutionary biology, but for a different purpose. He criticizes the standard assumptions of utility functions in game theory, in particular the *homo economicus* model. Among the questions he asks are: why are people willing to work for lower wages than they can earn elsewhere? Or, putting it differently, why do people deviate from what we would expect under standard economic assumptions, namely maximizing individual economic payoffs? He explains this in terms of non-economic motivations, such as people finding their current jobs more interesting or more useful to society, or feeling that their relationships with colleagues substitute for the family.

Vasin also examines whether it is possible to endogenize utility functions and identify how they evolve. Evolutionary game theory indicates that evolutionary stable strategies in self-reproducing populations maximize the fitness of individuals and, as a result, also the individual reproduction rate. Like Stark and Y.Q. Wang, Vasin focuses on altruistic and cooperative behaviour between relatives. He shows that such behaviour is evolutionary stable if it maximizes the total fitness of the family. However, he points out that in both human and non-human populations there are factors that limit the prevalence of altruistic and cooperative behaviour. In particular this behaviour is not protected from the invasion of selfish agents.

In any case, in his opinion, demographic data show that modern human populations maximize neither individual, nor family, nor population fitness. In fact, he suggests that the typical payoff functions of individuals are based on auxiliary utility functions (affected by feelings of pleasure and by consumerism) that maximize the fitness of 'superindividuals' (corporations,

organizations and institutions) who use a given human population as a resource for reproduction, and, to this end, manipulate people's utility functions.

In the next two chapters – by Englmaier and Chillemi respectively – the impact of altruism and social preferences is examined from a somewhat different, but related, angle, and in a different setting – that of labour markets and industrial relations. A central question facing economists has been how to design the right incentives to ensure that workers perform as the Principal desires. Englmaier in his chapter, 'Moral Hazard, Contracts and Social Preferences' provides a survey of recent contributions in the emerging field of behavioural contract theory that try to incorporate social preferences into the analysis of optimal contracts in situations of moral hazard.

Real-world contracts seldom follow economists' theoretical predictions that are based on the assumption that the agent is entirely self-interested. This view misses out on important factors which affect people's workplace choices and the contracts they sign – in particular their social preferences. Social ties in the workplace, altruistic relations with co-workers, team spirit and work morale, ideas about fairness (which play out differently in relation to fellow workers as versus employers), all matter. For instance, patterns of reciprocity and notions of fairness play an important role in human interactions and especially in labour markets, where people work closely together. People might care not only about their own payoff but also how payoffs are distributed amongst their fellow workers. They may have a social preference for equality among co-workers and might rather forego profits than accept inequitable distributions, since inequality causes them disutility (that is, they may have inequality aversion).[7] The preferences of agents can also exhibit inequality aversion when they compare themselves to the Principal. Agents may suffer a utility loss if they fail to get their fair share of the output, if the allocation is seen as being 'inequitable'. Taking such social preferences into account can explain behaviour that would appear irrational within the standard economic framework.

Basically, the survey shows how incorporating social preferences in economic models can enhance our understanding of relationships in the industrial workplace and add valuable insights to the analysis of incentive provisions. It also shows how these social preferences can be modelled. These aspects are examined both for the standard one-agent-one-principal problem *and* for multiple-agent settings and team production problems. In these models, a utility function is specified such that a separable term is added to standard utility derived from one's own income, to capture the disutility experienced when others get unequal incomes (that is, to capture relative income comparisons). Recently, experimental and field evidence has also helped amend standard utility functions to take account of social preferences.

Overall the survey shows that social preferences interact in non-trivial ways with incentives and alter the structure of optimal compensation schemes, sometimes drastically. So far the results are inconclusive with regard to the question: under what circumstances is a fair-minded workforce desirable (from the employer's viewpoint). The insights gained from choosing an optimal structure of incentives are still ambiguous, even in the settings of the tournament and team models worked out in this literature. According to Englmaier, so far the main advantage of these new research contributions is in their opening the door to a fruitful dialogue with researchers in the field of human resource management. This, he argues, can provide a promising avenue for future research. Related issues that he also highlights as worthy of further investigation are the implications of social preferences for structuring work teams, the production process and the information environment.

Chillemi's chapter – 'Mutual Concern, Workplace Relationships and Pay Scales' – also focuses on the impact of altruism in the work place. He examines how altruism among co-workers affects the performance of effort-enhancing incentive schemes and the firm's profits, based on Rotemberg's notion of trusty altruism. He outlines Rotemberg's investigation into whether friendly relations in the workplace can induce altruistic feelings among co-workers, thus helping to solve the free-rider problem in team production. Each worker chooses his degree of altruism with the intent of maximizing his own material surplus. Chillemi notes that crucial to Rotemberg's results is good fellowship that allows each worker to recognize the true attitude of his fellow workers, and also makes commitment possible.

Drawing on Rotemberg's work, Chillemi outlines a model to explain the fact that firms rarely adopt pay schemes based on worker competition. This model is then used to characterize the most profitable incentive scheme for maximizing the workers' efforts. Chillemi finds that under reasonable circumstances the firm's surplus increases with worker altruism. An interesting issue discussed in this chapter is whether altruism is consistent with (substantive) rationality. More specifically do workers choose their altruism parameters in order to maximize their material surplus? When workers choose these parameters cooperatively a strictly positive level of altruism emerges in equilibrium. In the case of endogenous altruism, becoming altruistic does not appear to be the best choice, but the scheme of incentives can be modified so that both the Principal and the workers gain.

The six chapters in the first part of this volume thus interweave psychology and economics theoretically, to challenge many of the assumptions and formulations of standard economics. The four chapters which follow in the second part of the volume carry forward this interdisciplinary exchange between psychology, economics and the cognitive sciences by using laboratory and field experiments, again to broaden our notions of rationality, and to take account of altruism as an important constituent of human motivation.

Part II Laboratory and field experiments

Fehr and Tyran in their chapter on 'Expectations and the Effects of Money Illusion', examine the nexus between cognitive psychology and rationality also discussed by Vercelli in Part I. In particular, they focus on the effects of money illusion. The authors argue that until recently money illusion was anathema to macroeconomists who tended to dismiss the 'psychological' explanation especially for two reasons: first because it contradicted the basic assumption of (narrow) rationality in economics (the argument being that rational human beings do not exhibit illusions, and if by assumption people behave rationally, there is little to study!). On this ground, money illusion stands rejected *a priori*. Second, macroeconomists rejected money illusion on the grounds that it was neither relevant nor backed by convincing evidence. It was seen as irrelevant on the argument that those suffering from such an illusion would lose economically, and this would provide a strong incentive to take illusion-free decisions. Fehr and Tyran emphasize that this argument is seriously flawed since it neglects the indirect effects of money illusion in shaping expectations, even if the individual-level effects are small and transitory.

The authors design experiments to investigate whether money illusion causes nominal inertia. Their results show that money illusion can have massive aggregate effects under conditions of strategic complementarity. Two types of aggregate effects are demonstrated. First, the authors show that money illusion is the cause of nominal inertia after an anticipated monetary shock in an economy with a unique equilibrium. Second, money illusion can have permanent effects by coordinating individuals on inferior equilibria. The use of the experimental method also makes it possible to precisely identify the conditions under which rational expectation models are correct, and the conditions under which they fail to capture important economic facts and forces.

The results obtained by Fehr and Tyran are similar to those highlighted by Vercelli in his chapter through a different approach. When the environment is sufficiently complex, characterized by a multiplicity of equilibria and strategic complementarity, money illusion (understood as a deviation from the rational-expectations equilibrium) may have large and permanent effects, despite limited individual-level deviations. The findings from experiments thus corroborate those obtained via cognitive psychology through questionnaires, namely that money illusion has a ubiquitous framing effect since the nominal representation of economic processes is often simpler and more salient. A fully rational agent expects a certain degree of money illusion from at least some of the other agents, and this is enough to produce macroeconomic inertia, even if individual money illusion is minute or non-existent.

Laboratory experiments also provide insights in Kritikos and Bolle's chapter on 'Utility-Based Altruism'. The chapter focuses on what we had noted to be

one of the central themes of this volume, namely the existence and nature of altruism and its impact on economic behaviour. The authors argue that one of the most prominent experiments used by economists for testing the existence of altruism is the Dictator Game. Based on this game, Kritikos and Bolle seek to provide evidence on economic approaches to the study of altruism, where the recipient's information status is variable. The standard Dictator Game serves as their benchmark to which they compare a modified game, where the recipient has incomplete information about the size of the pie.

The authors confirm the relevance of altruism in understanding economic behaviour, but note that its features depend on environmental conditions and subjective framing. In particular they explore the influence of cognitive psychology on the existence and degree of altruism. To this end the experiments compare the willingness of 'dictators' to make more or less altruistic offers to an anonymous recipient whose information status varies. The experiments show that it matters what degree of information the 'dictator' attributes to the recipient. This cannot be explained by the usual income-based approach favoured in economics, but can be explained easily by the utility-based approach favoured in psychology (and first applied to this problem in economics by Gary Becker), provided that a fairness component is built into the utility function.

Jungeilges and Theisen in their chapter, 'Equity Judgements Elicited through Experiments', also generate data from experiments in order to examine whether individual decisions are consistent with the Rawlsian second principle of justice.[8] According to the utilitarian school, welfare judgments should be based on how policies affect the sum of individual utilities, but according to the Rawlsian school welfare judgments should be based on how policies affect the utility of the worst-off individual in society.

Earlier empirical research has indicated that actual choices are determined by a mix of ethical and selfish considerations that are context-dependent, since they are affected by the constraints under which choices are made, and by strategic factors. In order to recover the underlying ethical principles from observed choices, Jungeilges and Theisen use experiments for eliciting the principles that guide individuals, when prioritizing on behalf of society. Specifically they test whether or not individuals make decisions in accordance with Rawls' second principle. They do so by asking students their responses to six different contexts of choice, each of which has a distributional consequence. The subjects are also asked about their demographic characteristics, parental employment background, and so on, and are selected from among business administration students at two stages of their education.

The authors apply sophisticated statistical and econometric techniques to the data generated by their experiments in order to extract the maximum possible information in the most reliable way. Their results show that the support for Rawlsianism declines with changes in a single factor in each choice situation. Utilitarian logic could be used to explain such a decline, but this is by

no means the only possible rationale. Equity judgments typically influence decisions according to the background of the decision maker. In the binary response models worked out by the authors, gender, parental background and education have a statistically significant effect. They find that women's behaviour is closer to Rawlsian principles than men's behaviour, and that subjects who list their parental background as self-employment tend to display a more selfish attitude than those with other parental backgrounds. However, the effect of different educational levels, while found to be relevant, is unclear and needs to be explored further.

The last chapter in this volume – 'Groups, Commons and Regulations' – by Juan Camilo Cardenas, uses both laboratory and field experiments to examine the impact of regulation, in a situation characterized by a negative externality due to the excessive exploitation of a natural resource. Most experimental studies are done in classrooms and in developed-country contexts. Cardenas departs substantially from this in conducting comparative experiments both in the field and in the classroom, in 10 different sites across rural Columbia. The field experiments involve villagers who have joint access to the natural resource. The focus of Cardenas' experimental research is also on the less-studied coordination problem, namely the management of common pool resources, critical for shedding light on questions of environmental sustainability.

The chapter explores the choices individuals make when an external regulation is introduced to solve the coordination problem. It examines how individuals vote when asked their preference regarding the application of such regulations by an external regulator, such as the state. The results suggest that even if a majority of players vote against the externally-imposed regulations, they are still willing to cooperate and reduce over-extraction. However, the players do not respond substantially to changes in the penalty size. Indeed players seem to cooperate 'too much' under a low penalty, and free-ride 'too much' under a high penalty. The results confirm that neither students nor villagers take decisions according to the canonical *homo economicus* model, nor achieve the socially optimum condition. These deviations from standard economic theory may be accounted for only by considering psychological and cultural factors that the author has sought to elicit through personal interviews.

A comparison between the responses of students and villagers is also of interest. Cardenas finds similarities on some counts and differences on others. On the differences, for instance, the villagers are found to be more opposed to external regulatory interventions than the students and to reject such regulations more often than the students. But the villagers are more inclined to cooperate under a non-binding setting. If the experiments had been conducted with students alone, without replication in the field with subjects who have practical familiarity with the problem, the research would have missed relevant information and provided fewer insights. The behavioural

differences found between college students and villagers would not only be of general interest to those using experimental methods, but would also caution against light-heartedly extending experimental results obtained with students, using the standard experimental economics method, to the population at large, or to different sections of it.

* * *

In sum, the chapters in this book contribute in varied ways to an important and innovative stream of research in economics, and we hope our readers will find them stimulating. As the chapters show, economics can considerably extend its theoretical and empirical scope by incorporating insights from psychology and other disciplines and challenging standard economic assumptions. And therein is likely to lie both the continued relevance of economics and its long-term evolution.

Finally, we thank all the authors of this volume for their rich contributions and for their patient receptivity to our suggestions for revision. And we owe a very special thanks to Michael Kaser, IEA General Editor, who has been the hidden third in our selection of chapters, in our pursuit of authors, and in keeping us on schedule.

Notes

1 See also Rabin's (1998) review of psychological findings that are of particular relevance to economics.
2 On the potential effect of emotions on economic behaviour, see especially Elster (1998), who also emphasizes the dearth of work on how emotions actually influence behaviour.
3 The term 'money illusion' has been used by scholars in different ways, but broadly it relates to the tendency to think in terms of nominal rather than real monetary values.
4 See e.g. the review by Loewenstein (1992); see also Thaler (2000).
5 Interested readers might also see Rilling *et al*. (2002) who use the iterated prisoner's dilemma game in an experiment to investigate the neurobiological basis of cooperative social behaviour, such as reciprocal altruism.
6 See especially, Ostrom *et al*. (1994), and the survey in Baland and Platteau (1996).
7 On this, see also, Englmaier and Wambach (2002). Among other things, they show that inequity aversion among agents provides a plausible explanation for the predominance of linear wage schemes in real labour markets.
8 Rawls (1997: 302) spells out his second principle as below: 'Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity.' Jungeilges and Theisen's chapter focuses on (a).

References

Baland, J.-M. and J.F. Platteau (1996) *Halting Degradation of Natural Resources: Is there a Role for Rural Communities*? (Oxford and New York: Oxford University Press).

Dawkins, R. (1990) *The Selfish Gene*, 2nd edn (Oxford: Oxford University Press).

Elster, J. (1998) 'Emotions and Economic Theory', *Journal of Economic Literature*, vol. 36(1), pp. 47–74.

Englmaier, F. and A. Wambach (2002) 'Contracts and Inequity Aversion', CESifo Working Paper no. 809, University of Munich.

Falk, I and O. Stark (2001) 'Dynasties and Destiny: On the Roles of Altruism and Impatience in the Evolution of Consumption and Bequests', *Economica*, vol. 68, pp. 505–18.

Glimcher, P.W. (2003) *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics* (Cambridge, MA: MIT Press).

Loewenstein, G. (1992) 'The Fall and Rise of Psychological Explanations in the Economics of Intertemporal Choice', in G. Loewenstein, and J. Elster (eds), *Choice over Time* (New York: Russell Sage Foundation), pp. 3–34.

Ostrom, E., J. Walker and R. Gardner (1994) *Rules, Games and Common-Pool Resources* (Ann Arbor MI: University of Michigan Press).

Rabin, M. (1998) 'Psychology and Economics', *Journal of Economic Literature*, vol. 36(1), pp. 11–46.

Rawls, J. (1997) *A Theory of Justice* (Cambridge, MA: The Belknap Press of Harvard University Press), twenty-second printing.

Rilling, J.K., D.A. Gutman, T.R. Zeh, G. Pagnoni, G.S. Berns and C.D. Kitts (2002) 'A Neural Basis for Social Cooperation', *Neuron*, vol. 35, pp. 395–405.

Sen, A.K. (1977) 'Rational Fools: A Critique of the Behavioural Foundations of Economic Theory', *Philosophy and Public Affairs*, vol. 6, pp. 317–44.

Sen, A.K. (2002) *Rationality and Freedom* (Cambridge MA: Harvard University Press).

Simon, H.A. (1982) 'From Substantive to Procedural Rationality', in H.A. Simon, *Models of Bounded Rationality* (Cambridge, MA: MIT Press), vol. 2, pp. 424–43.

Smith, A. [1759] (1966) *The Theory of Moral Sentiments*, Reprints of Economic Classics (New York: Augustus M. Kelley).

Stark, O. (1995) *Altruism and Beyond: An Economic Analysis of Transfers and Exchanges Within Families and Groups* (Cambridge: Cambridge University Press).

Thaler, R. (2000) 'From Homo Economicus to Homo Sapiens', *Journal of Economic Perspectives*, vol. 14(1), pp. 133–41.

This page intentionally left blank

Part I

Analytical Models and Methodological Issues

This page intentionally left blank

2

Self-Confidence and Personal Motivation*

Roland Bénabou
Woodrow Wilson School, Princeton University, USA
and
Jean Tirole
Institute of Industrial Economics, University of Toulouse, France

'Believe what is in the line of your needs, for only by such belief is the need fulfilled... Have faith that you can successfully make it, and your feet are nerved to its accomplishment.'
(William James, *Principles of Psychology*)

'I have done this, says my memory. I cannot have done that, says my pride, remaining inexorable. Finally–memory yields.'
(Friedrich Nietzsche, *Beyond Good and Evil*)

'I had during many years followed the Golden Rule, namely, that whenever a published fact, a new observation or thought came across me, which was opposed to my general results, to make a memorandum of it without jail and at once; for I had found by experience that such (contrary and thus unwelcome) facts and thoughts were far more apt to escape from memory than favourable ones.'
(Charles Darwin in *The Life of Charles Darwin*, by Francis Darwin)

* Reprinted with the permission of the MIT Press from *Quarterly Journal of Economics* (2002), vol. 117(3), pp. 871–915. We are grateful for helpful comments and discussions to Dilip Abreu, Olivier Blanchard, Isabelle Brocas, Ed Glaeser, Dan Gilbert, Ian Jewitt, David Laibson, George Loewenstein, Andrew Postlewaite, Marek Pycia, Matt Rabin, Julio Rotemberg and three anonymous referees, as well as conference and seminar participants at Chicago, Columbia, Cornell, MIT, the NBER, Northwestern, NYU, the Oxford Young Economists' Conference, the University of Pennsylvania, Princeton, Stanford, and Yale. Roland Bénabou gratefully acknowledges financial support from the National Science Foundation (SES-0096431).

1 Introduction

The maintenance and enhancement of self-esteem has always been identified as a fundamental human impulse. Philosophers, writers, educators and of course psychologists have all emphasized the crucial role played by self-image in motivation, affect and social interactions. The aim of this chapter is to bring these concerns into the realm of economic analysis, and show that this has important implications for how agents process information and make decisions. Conversely, the tools of economic modelling can help shed light on a number of apparently irrational behaviours documented by psychologists.

Indeed, both the demand and the supply sides of self-confidence appear at odds with economists' view of human behaviour and cognition. Why should people prefer rosy views of themselves to accurate ones, or want to impart such beliefs to their children? From car accidents, failed dot.com firms and day-trading to the space-shuttle disaster and lost wars, the costs of overconfidence are plain for all to see. Even granting that some 'positive illusions' could be desirable, is it even possible for a rational, Bayesian individual to deceive himself into holding them? Finally, the welfare consequences of so-called self-serving beliefs are far from clear: while 'thinking positive' is often viewed as a good thing, self-deception is not, even though the former is only a particular form of the latter.

To analyse these issues, we develop a simple formal framework that unifies a number of themes from the psychology literature, and brings to light some of their economic implications. We first consider the demand side of self-confidence, and identify in section 2 three main reasons why people may prefer optimistic self-views to accurate ones: a consumption value, a signalling value, and a motivation value. First, people may just derive utility from thinking well of themselves, and conversely find a poor self-image painful. Second, believing – rightly or wrongly – that one possesses certain qualities may make it easier to convince others of it. Finally, confidence in their abilities and efficacy can help individuals undertake more ambitious goals and persist in the face of adversity. While we shall mostly focus on this last explanation, all three should be seen as complementary, and for many purposes work equally well with the supply side of our model (self-deception).

The main reason why we emphasize the motivation theory is its substantially broader explanatory power. Indeed, it yields an endogenous value of self-confidence that responds to the situations and incentives which the individual faces, in a way that can account for both 'can-do' optimism and 'defensive' pessimism. It also readily extends to economic and social interactions (altruistic or not), explaining why people generally prefer self-confident partners to self-doubting ones, and invest both time and effort in supporting the latter's morale.

The first premise of the motivation theory is that people have imperfect knowledge of their own abilities, or more generally of the eventual costs and payoffs of their actions.[1] The second is that ability and effort interact in determining performance; in most instances they are complements, so that *a higher self-confidence enhances the motivation to act*. As demonstrated by the opening quotation from James (1890), this complementarity has long been familiar in psychology.[2] It is also consistent with the standard observation that morale plays a key role in difficult endeavours; conversely, when people expect to fail they fail quite effectively, and failure leads to failure more readily for individuals characterized with low self-esteem (Salancik, 1977).

The fact that a higher self-confidence enhances the individual's motivation gives anyone with a vested interest in his performance an incentive to build up and maintain his self-esteem. First, the manipulator could be another person (parent, teacher, spouse, friend, colleague, manager) who is eager to see him 'get his act together', or otherwise apply himself to the task at hand. Such *interpersonal strategies* are studied in Bénabou and Tirole (2003). Second, for an individual suffering from time inconsistency (for example hyperbolic discounting), the current self has a vested interest in the self-confidence of future selves, as it helps counter their natural tendency to quit too easily. It is in this context, which builds on Carrillo and Mariotti (2000), that we shall investigate a variety of *intrapersonal strategies* of self-esteem maintenance. We shall thus see how and when people may choose to remain ignorant about their own abilities, and why they sometimes deliberately impair their own performance or choose overambitious tasks in which they are sure to fail (self-handicapping).

Section 3 thus turns to the supply side of the self-confidence problem, and the 'reality constraints' that limit the extent to which people can engage in wishful thinking. In our model we maintain the standard assumption of individuals as rational (Bayesian) information processors. While almost universal in economics, this view is more controversial in psychology. On one hand, a lot of the classical literature has emphasized rationality and information-seeking in the process of self-identification, documenting the ways in which people update their beliefs according to broadly Bayesian principles.[3] On the other hand, the more recent cognitive literature abundantly documents the less rational (or at least, subjectively motivated) side of human inference.

For instance, a substantial body of evidence suggests that people tend to recall their successes more than their failures, and have self-servingly biased recollections and interpretations of their past performances.[4] Similarly, they tend to overestimate their abilities and other desirable traits, as well as the extent to which they have control over outcomes. They also rate their own probabilities as above average for favourable future life events, and below average for unfavourable ones; the more controllable these events through their future actions, the more so.[5]

We shall capture this class of *self-deception* phenomena with a simple game-theoretic model of endogenous memory, or awareness-management, which represents one of the main contributions of this chapter. Drawing on evidence about the mechanics and limitations of memory, it shows how to reconcile the motivated ('hot') and rational ('cold') features of human cognition, and could be used in any setting where a demand for motivated beliefs arises. The basic idea is that the individual can, within limits and possibly at a cost, *affect the probability of remembering* a given piece of data. At the same time, we maintain rational inference, so people realize (at least to some extent) that they have a selective memory or attention.

The resulting structure is that of a game of strategic communication between the individual's temporal selves. In deciding whether to try to repress bad news, the individual weighs the benefits from preserving his effort motivation against the risk of becoming overconfident. Later on, however, he appropriately discounts the reliability of rosy recollections and rationalizations. The implications of this game of asymmetric information are quite different from those of *ex ante* decisions about information acquisition (for example self-handicapping or selective search). In particular, multiple intrapersonal equilibria ('self-traps') may arise, ranging from systematic denial to complete self-honesty. More generally, we characterize the set of perfect Bayesian equilibria and its dependence on the individual's degree of time inconsistency and repression costs ('demand and supply' parameters).

The model also has interesting implications for the distribution of optimism and pessimism across agents, which we examine in section 4. We show that when the costs of repression are low enough, most people typically believe themselves to be more able than they actually are, as well as more able than both the average and the median of the population. A minority will have either realistically low assessments, or actually severely underestimate themselves. We also highlight the key role played by Bayesian-like introspection (understanding, at least partially, one's own incentives for self-esteem maintenance) in the model's results, and why incorporating this essential human trait is more fruitful than modelling agents as naively taking all recollections and self-justifications at face value.

Section 5 examines the welfare impact of equilibrium self-deception. Is a more active self-esteem maintenance strategy, when chosen, always beneficial? How can people be 'in denial' if it does not serve their best interests? We show that, in addition to the trade-off mentioned earlier between the confidence-maintenance motive and the risks of overconfidence, *ex ante* welfare reflects a third effect, namely the spoiling of good news by self-doubt. Intuitively, when adverse signals about his ability are systematically repressed, the individual can never be sure that only positive ones were received, even when this is actually true. We characterize the conditions under which always 'looking at the bright side' pays off on average or, conversely, when it would be better to always 'be honest with yourself', as Charles Darwin apparently concluded.

In section 6 we turn to the case where ability and effort are substitutes rather than complements. This typically occurs when the payoff for success is of a 'pass–fail' nature, or characterized by some other form of satiation. Since a high perceived ability may now increase the temptation to exert low effort ('coasting'), this case allows us to account for what psychologists refer to as 'defensive pessimism': the fact that people sometimes minimize, rather than aggrandize, their previous accomplishments and expectations of future success. Another variant of the model considered in this section involves replacing the motivation value of self-confidence by a purely affective one. Section 7 concludes the chapter. All proofs are gathered in the Appendix.

This chapter is related to several strands of the new literature that tries to better link economics and psychology. A hedonic concern for self-image, in the form of preferences over beliefs, was first explored in Akerlof and Dickens' (1982) well-known model of dissonance reduction, and more recently in Rabin (1995), Weinberg (1999) and Köszegi (1999). In emphasizing an endogenous value of self-confidence and retaining the constraint of Bayesian rationality, our chapter is most closely related to the work of Carrillo and Mariotti (2000), who first showed how information manipulation may serve as a commitment device for time-inconsistent individuals (see also Brocas and Carrillo, 1994). The central role played by memory also relates our model to those of Mullainathan (2002) and Laibson (2001), although one of its main features is to make recall endogenous.

2 The demand for self-confidence

In most societies, self-confidence is widely regarded as a valuable individual asset. Going back at least to William James, an important strand in psychology has advocated 'believing in oneself' as a key to personal success. Today, an enormous 'self-help' industry flourishes, a sizable part of which purports to help people improve their self-esteem, shed 'learned helplessness' and reap the benefits of 'learned optimism'.[6] American schools place such a strong emphasis on imbuing children with self-confidence ('doing a great job!') that they are often criticized for giving it preeminence over the transmission of actual knowledge. Hence the general question: why is a positive view of oneself, as opposed to a fully accurate one, seen as such a good thing to have?

Consumption value. A first reason may be that thinking of oneself favourably just makes a person happier: self-image is then simply another argument in the utility function. Indeed, psychologists emphasize the affective benefits of self-esteem as well as the functional ones on which we shall focus. One may also hypothesize that such preferences over beliefs could have been selected through evolution: the overconfidence that typically results may propel individuals to undertake activities (exploration, foraging, combat) which are more risky than warranted by their private material returns, but confer important external benefits on the species. In section 5 we shall

explain how a hedonic self-image motive can readily be incorporated into our general framework.

Signalling value. A second explanation may be that believing oneself to be of high ability or morality makes it easier to convince others (rightly or wrongly) that one does have such qualities. Indeed, it is often said that to lie most convincingly a person must believe his own lies. While the idea that people are 'transparent' and have trouble misrepresenting their private information may seem unusual in economics, one could easily obtain an instrumental value of self-confidence from a signalling game where those who truly believe in their own abilities face lower costs of representing themselves favourably to others.

Motivation value. The explanation that we emphasize most is that self-confidence is valuable because it improves the individual's motivation to undertake projects and persevere in the pursuit of his goals, in spite of the setbacks and temptations that periodically test his willpower. Morale is universally recognized as key to winning a medal, performing on stage, getting into college, writing a great book, doing innovative research, setting up a firm, losing weight, finding a mate and so forth. The link between self-confidence and motivation is also pervasive in the psychology literature, from early writers like James (1890) to contemporary ones like Bandura (1977), according to whom '*beliefs of personal efficacy constitute the key factor of human agency*' (see also, e.g., Deci, 1975, or Seligman, 1990). The motivation theory also readily extends to economic (non altruistic) interactions, explaining why people typically prefer self-confident co-workers, managers, employees, team-mates, soldiers, and so on, to self-doubting ones; and why they spend substantial time and effort supporting the morale of those with whom they end up being matched.[7]

The motivation problem

> 'Had I been less definitively determined to start working, I might have made an effort to begin right away. But because my resolve was absolute and, within twenty-four hours, in the empty frames of the next day where everything fit so well since I was not yet there, my good resolutions would easily be accomplished, it was better not to choose an evening where I was ill disposed for a beginning to which, alas! the following days would turn out to be no more propitious.'
>
> (Marcel Proust, *Remembrance of Things Past*)

Consider a risk-neutral individual with a relevant horizon of three periods: $t = 0, 1, 2$. At date 0, he selects an action that potentially affects both his flow payoff u_0 and his date-1 information structure.[8] At date 1, he decides whether to undertake a task or project (exert effort, which has disutility cost $c > 0$) or not (exert no effort). With some probability θ, which defines his *ability*, the project will succeed and yield a benefit V at date 2; failure generates

no benefit. The individual's beliefs over θ (defining his *self-confidence* or self-esteem) are described by distribution functions $F(\theta)$ at date 0 and $F_1(\theta)$ at date 1. In the intervening period new information may be received, or previous signals forgotten; we shall focus here on the first, more standard case, and turn to memory in section 3. Note that with risk-neutrality the mean $\bar{\theta}_1 \equiv \int_0^1 \theta dF_1(\theta)$ will be a sufficient statistic for F_1; for brevity we shall also refer to it as the agent's date-1 self-confidence.

Finally, we assume that the individual's preferences exhibit *time-inconsistency*, due to quasi-hyperbolic discounting. There is indeed considerable experimental and everyday evidence that intertemporal choices exhibit a 'salience of the present', in the sense that discount rates are much lower at short horizons than at more distant ones.[9] Denoting u_t and $\mathsf{E}_t[\cdot]$ the flow payoffs and expectations at $t = 0, 1, 2$, the intertemporal utility perceived by the individual as of date 1 is:

$$u_1 + \beta\delta \mathsf{E}_1[u_2] = -c + \beta\delta\bar{\theta}_1 V \tag{1}$$

when he undertakes the activity, and 0 when he does not. By contrast, intertemporal utility conditional on the same information set at date 1, but evaluated from the point of view of date zero is:

$$u_0 + \beta \mathsf{E}_0[\delta u_1 + \delta^2 u_2 \,|\, \bar{\theta}_1] = u_0 + \beta\delta[-c + \delta\bar{\theta}_1 V] \tag{2}$$

if the activity is undertaken at date 1, and u_0 otherwise.[10] Whereas δ is a standard discount factor, β reflects the momentary salience of the present. When $\beta < 1$ the individual at date 0 ('Self 0') is concerned about his date 1 ('Self 1's') excessive preference for the present, or *lack of willpower*, which leads to the underprovision of effort (procrastination). Indeed, Self 1 only exerts effort in the events where $\bar{\theta}_1 > c/\beta\delta V$, whereas, from the point of view of Self 0, it should be undertaken whenever $\bar{\theta}_1 > c/\delta V$. Note that while we focus here on the case where the individual's intrinsic ability θ is unknown, it could equally be the expected payoff in case of success V, the 'survival' probability δ, or the task's difficulty, measured by the cost of effort c. All that matters for our theory is that the individual be uncertain of the long term *return to effort* $\theta\delta V/c$ which he faces.

Confidence maintenance versus overconfidence

In an important paper, Carrillo and Mariotti (2000) showed that, in the presence of time inconsistency (TI), Blackwell garblings of information may increase the current self's payoff. This result can be usefully applied, and further developed, in our context.

Suppose that, at date zero, our individual can choose between just two information structures for date 1. In the finer one, Self 1 learns his ability θ exactly. In the coarser one, he learns nothing that Self 0 did not

know: $F_1(\theta) = F(\theta)$, and hence $\bar{\theta}_1 = \int_0^1 \theta dF_1(\theta) \equiv \bar{\theta}_F$. Let us first assume that, in the absence of information, Self 1 will undertake the task: $\bar{\theta}_F > c/\beta\delta V$. The value attached by Self 0 to Self 1's learning the value of θ is therefore $\beta\delta$ times

$$\mathcal{I}_F \equiv \int_{c/\beta\delta V}^{1} (\delta\theta V - c)dF(\theta) - (\delta\bar{\theta}_F V - c) = \mathcal{G}_F - \mathcal{L}_F, \text{ where} \tag{3}$$

$$\mathcal{G}_F \equiv \int_{0}^{c/\delta V} (c - \delta\theta V)dF(\theta) \tag{4}$$

$$\mathcal{L}_F \equiv \int_{c/\delta V}^{c/\beta\delta V} (\delta\theta V - c)dF(\theta) \tag{5}$$

$\mathcal{G}_F$ stands for the gain from being informed, which arises from the fact that better information reduces the risk of *overconfidence* on the part of Self 1. Overconfidence occurs when the individual's ability is below $c/\delta V$ but he is unaware of it, and thus inappropriately undertakes or perseveres in the project. $\mathcal{L}_F$ stands for the loss from being informed, which may depress the individual's self-confidence: if he learns that θ is in some intermediate range, $c/\delta V < \theta < c/\beta\delta V$, he will procrastinate at date 1 even though, *ex ante*, it was optimal to work. Information is thus detrimental to the extent that it creates a risk that the individual will fall into this time-inconsistency (TI) region. If this *confidence maintenance* motive is strong enough ($\mathcal{L}_F > \mathcal{G}_F$), the individual will prefer to remain uninformed: $\mathcal{I}_F < 0$. More generally, note that $\mathcal{I}_F$ is lower, the lower is β. By contrast, in the absence of time inconsistency ($\beta = 1$) we have $\mathcal{L}_F = 0$, and thus $\mathcal{I}_F \geq 0$: in classical decision theory, information is always valuable.

The overconfidence effect calls for more information, confidence maintenance for less. This trade-off has been noted by empirical researchers. For instance, Leary and Downs (1995) summarize the literature by noting that: (a) 'persons with high self esteem perform better after an initial failure and are more likely to persevere in the face of obstacles'; (b) 'high self-esteem is not always functional in promoting task achievement. People with high self-esteem may demonstrate non-productive persistence at insoluble tasks, thereby undermining their effectiveness. They may also take excessive and unrealistic risks when their self-esteem is threatened'.

To understand the last statement, let us turn to the case where $\bar{\theta}_F < c/\beta\delta V$. Since Self 1 now always exerts (weakly) less effort than Self 0 would like him to, information can only help the individual restore his deficient motivation. Indeed,

$$\mathcal{I}_F = \int_{c/\beta\delta V}^{1} (\delta\theta V - c)dF(\theta) > 0 \tag{6}$$

Moreover, $\mathcal{I}_F$ is now *higher*, the lower is β. In such situations the individual will avidly seek feedback on his ability, and his choices of tasks and social

interactions will have the nature of 'gambles for resurrection' of his self-esteem.

Putting together the different cases, we see that the value of information is *not monotonic* with respect to initial self confidence. Indeed, for someone with confidence so low that $\bar{\theta}_F < c/\beta\delta V, \mathcal{I}_F$ is always positive and increasing with respect to (stochastic) increases in θ.[11] For an individual with $F(c/\delta V) = 0$ but $F(c/\beta\delta V) < 1, \mathcal{I}_F$ is always negative. Finally, for a person so self-assured that $F(c/\beta\delta V) = 0$, motivation is not a concern (as if β were equal to 1), but neither is overconfidence: $\mathcal{I}_F = \mathcal{G}_F = \mathcal{L}_F = 0$. Therefore, there must exist some intermediate range where $\mathcal{I}_F$ first declines and becomes negative, then increases back towards zero.

What types of people are most eager to maintain their self-confidence?

Let us now consider two individuals with different degrees of initial self-confidence, and ask which one is least receptive to information. We denote their prior distributions over abilities as $F(\theta)$ and $G(\theta)$, with densities $f(\theta), g(\theta)$ and means $\bar{\theta}_F, \bar{\theta}_G$. To make confidence-maintenance meaningful, let $\bar{\theta}_F > \bar{\theta}_G > c/\beta\delta V$. For comparing levels of self-confidence, however, just looking at expected abilities turns out not to be sufficient.

Definition 1 An individual with distribution F over ability θ has higher self-confidence than another one with distribution G if the likelihood ratio $f(\theta)/g(\theta)$ is increasing in θ.

Abstracting for the moment from any cost attached to learning or not learning the true ability, it is easy to see from (3) that $\mathcal{I}_F \geq 0$ if and only if

$$\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)} d\theta \geq \left(\frac{1-\beta}{\beta}\right)\left(\frac{c}{\delta V}\right) \tag{7}$$

The monotone likelihood ratio property (MLRP) implies that $F(\theta)/F(c/\beta\delta V) \leq G(\theta)/G(c/\beta\delta V)$ for all $\theta \leq c/\beta\delta V$. Therefore, the left-hand side of (7) is smaller under F than under G, meaning that the person with the more positive self-assessment will accept information about his ability for a smaller set of parameters. Intuitively, he has more to lose from information, and is therefore more insecure.

Proposition 1 If an individual prefers not to receive information in order to preserve his self-confidence, so will anyone with higher initial self-confidence: if $I_G < 0$ for some distribution G over θ, then $I_F < 0$ for any distribution F such that the likelihood ratio f/g is increasing.

Self-handicapping

A well-documented and puzzling phenomenon is that people sometimes create obstacles to their own performance.[12] In experiments, subjects with

fragile self-confidence opt to take performance-impairing drugs before an intelligence test. In real life, people withhold effort, prepare themselves inadequately, or drink alcohol before undertaking a task. They also set themselves overambitious goals, where they are almost sure to fail. Test anxiety and 'choking' under pressure are yet other common examples. Psychologists have long suggested that self-handicapping is often a self-esteem maintenance strategy (instinctive or deliberate), directed both at oneself and at others.[13]

To examine this question, consider an individual with prior beliefs $F(\theta)$, faced at date zero with a choice between an efficient action that (for simplicity) will fully reveal his ability, and an inefficient, 'self-handicapping' one that entails an expected cost $h_0(F) \geq 0$, but is totally uninformative about θ. Assuming that $\bar{\theta}_F > c/\beta\delta V$ as before, equation (7) immediately generalizes to show that he will self-handicap if and only if $-\beta\delta I_F \geq h_0(F)$, or:

$$\left(\frac{1-\beta}{\beta}\right) c - \delta V \left[\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)} d\theta\right] \geq \frac{h_0(F)}{\beta\delta F(c/\beta\delta V)} \tag{8}$$

Note, first, that multiplying (8) by $F(c/\beta\delta V)$ yields a decreasing function of β on the left-hand side. Therefore, people who are more concerned about sustaining motivation (more time-inconsistent) are more likely to self-handicap, and will chose to do so when the short-run costs of doing so are not too large. Next, let us compare individuals with different prior beliefs about themselves. As before, those who are initially more self-confident have more to lose from learning about their ability: by the MLRP, the left-hand side of (8) is larger than if F were replaced with G. However, the more self-confident are also *less likely* to receive bad news, and this reduces the return on 'investing in ignorance': the MLRP implies that $F(c/\beta\delta V) \leq G(c/\beta\delta V)$, which tends to make the right-hand side of (8) also larger under F than under G. Thus, in general one cannot conclude whether people with higher or lower self-confidence are more likely to self-handicap.[14] This ambiguity is fundamentally linked to the non-monotonicity noted earlier for the value of information: while the MLRP ensures that the *sign of* $\mathcal{I}_F$ varies monotonically with initial beliefs, the *absolute amount* does not. When self-handicapping costs are relatively small, however – which is often the case in experiments – the 'more to lose' effect identified in Proposition 1 will prevail.

It is interesting in this respect to note that psychologists have also not reached a firm conclusion on whether high or low self-confidence people are the most defensive of their egos, although there seems to be somewhat more evidence in favour of the first hypothesis. Thus Greenier *et al.* (1995) contrast 'humanistically oriented theories . . . according to which high self-esteem individuals' feelings of self-worth are built on solid foundations that do not require continual validation', with experimental research showing that 'high self-esteem individuals are the more likely to display self-serving attributions,

self-handicap to enhance the potentially positive implications of good performance, set inappropriately risky goals when ego-threatened, and actively create less fortunate others with whom they can compare favorably.'

3 The psychological immune system

> 'Just as there was in his study a chest of drawers which he managed never to look at, which he went out of his way not to encounter when walking in or out, because in one drawer was held the chrysanthemum which she had given him on the first evening . . . so there was in him a place which he never let his mind approach, imposing on it, if necessary, the detour of a long reasoning . . . it was there that lived the memories of happier days.'
>
> (Marcel Proust, *Remembrance of Things Past*)

We now turn to the supply side of the self-esteem problem. Given that a positive self-assessment may be desirable (whether for motivational, signalling or hedonic reasons), what are the means through which it can be achieved, or at least pursued?

Wired-in optimism. A first hypothesis could be that evolution has selected for a particular cognitive bias in humans, causing them to systematically and involuntarily underweigh adverse signals about themselves, and overweigh positive ones. This explanation is rather problematic: the extent of overconfidence or overoptimism varies both over time and across tasks, and a great many people actually suffer from underconfidence (the extreme case being depression). Furthermore, individuals often 'work' quite hard at defending their self-image when it is threatened, going though elaborate schemes of denial, self-justification, furniture-avoidance and the like.

Blissful ignorance. When self-confidence is valuable capital, it may be preferable to remain uninformed than to put it at risk by exposing oneself to new information. In particular, as seen in the previous section, *ex ante* strategic ignorance (Carrillo and Mariotti, 2000) or even self-handicapping may help a time-inconsistent individual safeguard his motivation. In a context of hedonic beliefs, workers in a hazardous job may not want to know about the exact risks involved (Akerlof and Dickens, 1982).

Self-deception. Most often, the relevant issue is not whether to seek or avoid information *ex ante* (before knowing what it will turn out to say), but how to deal with the good and especially the bad news concerning one's performances and abilities that life inevitably brings. This is where the mechanisms of defensive denial, repression, self-serving attributions and the like, so

prominently emphasized in psychology, come into play. We shall capture this class of phenomena with a simple game-theoretic model of endogenously selective memory.

Managing awareness: the role of memory

Psychologists, and before them writers and philosophers, have long documented people's universal tendency to deny, explain away and selectively forget ego-threatening information. Freudian repression is the most obvious example, but various other forms of *motivated cognition* and *self-deception* feature prominently in contemporary psychology. Thus, a lot of research has confirmed that people tend to recall their successes more than their failures (e.g., Korner, 1950; Mischel *et al*., 1976), have self-servingly biased recollections of their past performances (Crary, 1996), and readily find 'evidence' in their personal histories that they possess characteristics which they view (sometimes as the result of experimental manipulation) as correlated with success in professional or personal life (Kunda and Sanitioso, 1989; Murray and Holmes, 1993). Similarly, they often engage in 'beneffactance', viewing themselves as instrumental for good but not bad outcomes (Zuckerman, 1979). When they commit a bad deed they reframe the facts to try and convince themselves that it was not so bad ('he deserved it', 'the damage was limited'), or attribute the responsibility to others (Snyder, 1985).

At the same time, the impossibility of simply choosing the beliefs we like has always stood in the way of a fully consistent theory of self-deception. Sartre (1953) argued that the individual must simultaneously know and not know the same information. Gur and Sackeim (1979) defined self-deception as a situation in which (a) the individual holds two contradictory beliefs; (b) he is not aware of holding one of the beliefs; and (c) this lack of awareness is motivated.

Our intertemporal model allows us to unbundle the 'self that knows' from the 'self that doesn't know', and thereby reconcile the motivation ('hot') and cognition ('cold') aspects of self-deception within a standard information-theoretic framework. The basic idea is that the individual can, within limits, *affect the probability of remembering* a given piece of data. Under time-inconsistency, there is an incentive to try to recall signals that help sustain long-term goals, and forget those that undermine them. This is the motivation part.[15] On the other hand, we maintain the rational inference postulate, so people realize (at least to some extent) that they have a selective memory or attention. This is the cognition part.

Assumption 1 (memory or awareness management). The individual can, at a cost, increase or decrease the probability of remembering an event or its interpretation.

Formally, let $\lambda \in [0, 1]$ denote the probability that a given piece of information received at date 0 will be recalled or accessed at date 1. We define the natural rate of recall $\lambda_N \in (0, 1]$ as that which maximizes the date 0 flow payoff u_0. Increasing or decreasing λ thus involves a 'memory cost' $M(\lambda)$, i.e. a reduction in u_0, with $M(\lambda_N) = 0, M'(\lambda) \leq 0$ for $\lambda < \lambda_N$ and $M'(\lambda) \geq 0$ for $\lambda > \lambda_N$.[16]

Whether it refers to the subconscious or points to the differential accessibility and decay of memories stored in specialized parts of the brain, virtually all modern psychology recognizes that only part of the individual's accumulated stock of information is readily available for conscious, purposive processing and decision-making. Furthermore, the encoding and retrieval process is subject to systematic influences, both internal and external: (a) information that is rehearsed often is better remembered (indeed, that is why we cram for an exam); conversely, if one is preoccupied or distracted when an event unfolds, one has greater difficulty remembering the details; and (b) direct behavioural experience makes the information more accessible in memory, because later on recall is more likely to be activated by situational *cues*.[17]

Such mechanisms seem to be at work in experiments where subjects who are asked to behave in a self-deprecating manner later report lower self-esteem than earlier, while persons who are asked to display self-enhancing behaviour report higher self-esteem (Jones *et al.*, 1981). This may be due to the fact that they were led to rehearse unfavourable or favourable information about themselves, thus increasing the probability of remembering it later on. Similarly, receiving positive feedback seems to trigger a cue-based 'warm-glow' effect, which automatically makes accessible to the individual other instances of himself in positive situations (Greenwald, 1980).

These frictions in the mechanics of memory give the individual some discretion about what data he is more likely to consciously recall later on – thereby opening the door to motivated cognition. Thus, a person who wishes to remember good news and forget bad news can linger over praise or positive feedback, rehearse it periodically, and choose to be more frequently in environments or with people who will remind him of his past successes.[18] Conversely, he can eschew situations that remind him of bad news, tear up the picture of a former lover, or, like the narrator in Proust's novel, avoid passing by a chest of drawers which contains cues to painful memories. He can work unusually hard to 'forget' (really, not think about) a failed relationship or family problem, or even use drugs and alcohol.

The individual can also find ways to discount self-threatening news in the first place. A common such strategy is to seek out information that derogates the informativeness of the initial data (Frey, 1981; Gilovich, 1991). After being criticized in a seminar or referee report, a researcher will look hard for reasons why the commenter has poor taste, a limited understanding of the issues, a vested interest in a competing theory or body of empirical evidence and so forth. Interpersonal strategies can also be called upon; thus, a verbal fight with one's spouse or someone who criticizes one's work may

(consciously or not) serve the purpose of creating a distraction that will impair accurate recollection of the details of the criticism (of course, it has costs as well . . .).

As these examples make clear, it is important to note that we need not *literally* assume that the individual can directly and mechanically suppress memories. Our model is equally consistent with a Freudian view where memories get buried in the unconscious (with some probability of reappearance), and with the more recent cognitive psychology which holds that memory itself cannot be controlled, but emphasizes the different ways in which *awareness* can be affected: the choice of attention when the information accrues, the search for or avoidance of cues, the process of selective rehearsal afterwards and again the choice of attention at the time the information is (voluntarily or accidentally) retrieved.[19] We shall therefore use the terms 'memory' and 'awareness of past informations' interchangeably.

> **Assumption 2** (metacognition). While the individual can manipulate his conscious self-knowledge, he is aware that incentives exist that result in selective memory.

As illustrated by the opening quotations from Nietzsche and Darwin, if a person has a systematic tendency to forget, distort or repress certain types of information he will likely become (or be made) aware of it, and not blindly take at face value the fact that most of what comes to his mind when thinking about his past performances and the feedback he received is good news. Instead, using (some) rational inference, he will realize that what he may have forgotten are not random events.[20] Formally, this *introspection* or skepticism with respect to the reliability of one's own self-knowledge is represented by Bayes' rule, which implies that a person cannot consistently fool himself in the same direction. Less sophisticated inference processes lead to similar results, so long as they are not excessively naive (see p. 40).

The game of self-deception

Let the agent receive, at date 0, a signal σ about his ability θ. To make things simple, let σ take only two values: with probability $1-q$ the agent receives bad news, $\sigma = L$, and with probability q he receives no news at all, $\sigma = \emptyset$. In other words, 'no news is good news'. Let

$$\theta_L \equiv \mathsf{E}[\theta \mid \sigma = L] < \mathsf{E}[\theta \mid \sigma = \emptyset] \equiv \theta_H \tag{9}$$

Since σ is informative about the return to date-1 effort, the agent's Self 1 would benefit from having this signal. If it is ego-threatening, however, Self 0 may have an interest in suppressing it. The recollection at date 1 of the news σ will be denoted $\hat{\sigma} \in \{\emptyset, L\}$. We assume that memories can be lost but not manufactured *ex nihilo*, so $\sigma = \emptyset$ always leads to $\hat{\sigma} = \emptyset$. A signal $\sigma = L$, on

the other hand, may be forgotten due to natural memory decay or voluntary repression. Let λ denote the probability that bad news will be remembered accurately:

$$\lambda \equiv \Pr[\hat{\sigma} = L \mid \sigma = L] \tag{10}$$

As explained earlier, the agent can increase or decrease this probability with respect to its 'natural' value $\lambda_N \leq 1$; choosing a recall probability λ involves a 'memory cost' $M(\lambda)$. We shall now analyse the equilibrium in several stages.

1. Inference problem of Self 1. Faced with a memory $\hat{\sigma} \in \{L, \emptyset\}$, Self 1 must first assess its credibility. Given that memories cannot be invented, unfavourable ones are always credible. When Self 1 does not recall any adverse signals, on the other hand, he must ask himself whether there was indeed no bad news at date 0, or whether it may have been lost or censored. If Self 1 thinks that bad news is recalled with probability λ^*, he uses Bayes' rule to compute the *reliability* of a 'no recollection' message as

$$r^* \equiv \Pr[\sigma = \emptyset \mid \hat{\sigma} = \emptyset; \lambda^*] = \frac{q}{q + (1-q)(1-\lambda^*)} \tag{11}$$

His degree of self-confidence is then

$$\theta(r^*) \equiv r^*\theta_H + (1 - r^*)\theta_L \tag{12}$$

2. Decisions and payoffs. We normalize the payoff in case of success to $V = 1$, and assume that the cost of date 1 effort is drawn from an interval $[\underline{c}, \bar{c}]$, with probability distribution $\Phi(c)$ and density $\varphi(c) > 0$. We assume that $\bar{c} > \beta\delta\theta_H > \beta\delta\theta_L > \underline{c}$, which means that at date 1 there is always a positive probability of no effort, and a positive probability of effort.[21]

Given a signal σ at date 0 and a memory $\hat{\sigma}$ at date 1, Selves 0 and 1 respectively assess the productivity of date 1 effort as $E[\theta \mid \hat{\sigma}]$ and $E[\theta \mid \hat{\sigma}]$. Self 1 only works when the realization of the effort cost is $c < \beta\delta E[\theta \mid \hat{\sigma}]$. so Self 0's payoff is

$$\beta\delta \int_0^{\beta\delta \, \mathsf{E}[\theta \mid \hat{\sigma}]} (\delta \, \mathsf{E}[\theta|\sigma] - c) d\Phi(c) \tag{13}$$

To build intuition, suppose for a moment that Self 0 could freely and costlessly manipulate Self 1's expectation, $E[\theta \mid \hat{\sigma}]$. What beliefs would he choose for a naive Self 1? Maximizing (13), we find that Self 0 would like to set

$E[\theta \mid \hat{\sigma}]$ equal to $E[\theta \mid \sigma]/\beta$. This makes clear how time consistency gives Self 0 an incentive to boost or maintain Self 1's self-confidence; the problem, of course, is that Self 1 is not so easily fooled. These two effects are consistent with the common view in psychology that a moderate amount of 'positive illusion' about oneself is optimal, but that many people find it quite difficult to strike this desirable balance.

3. Costs and benefits of selective memory or attention. Focusing on the 'bad news' case, denote as $U_C(\theta_L \mid r^*)$ the expected utility of Self 0 (gross of memory-management costs) when the adverse information is successfully forgotten, and as $U_T(\theta_L)$ the corresponding value when it is accurately recalled. The subscripts C and T stand for 'censored' and 'truth' respectively. Hiding from Self 1 the signal $\sigma = L$ raises his self-confidence from θ_L to $\theta(r^*)$, leading him to exert effort in the additional states of the world where $\beta\delta\theta_L < c < \beta\delta\theta(r^*)$. As with *ex ante* ignorance, this has both costs and benefits; thus if r^* is high enough that $\beta\theta(r^*) > \theta_L$, the *net gain or loss from self-deception* is

$$U_C(\theta_L \mid r^*) - U_T(\theta_L) = \beta\delta \left(\int_{\beta\delta\theta_L}^{\delta\theta_L} (\delta\theta_L - c) d\Phi(c) - \int_{\delta\theta_L}^{\beta\delta\theta(r^*)} (c - \delta\theta_L) d\Phi(c) \right) \tag{14}$$

The first integral is decreasing in β, becoming zero at $\beta = 1$: it represents the *gain from confidence-building*, which alleviates Self 1's motivation problem. The second integral is increasing in β: it reflects the *loss from overconfidence*, which causes Self 1 to attempt the task in states of the world where even Self 0 would prefer that he abstain. Note that these effects are now endogenous. Thus the overconfidence cost in (14) is larger, the more reliable Self 1 considers the memory process to be, that is the larger r^*. Conversely, if r^* is so low as to have $\beta\theta(r^*) < \theta_L$ the overconfidence effect disappears entirely, but the confidence-building effect is now limited by $\theta(r^*)$.

4. Strategic memory or awareness management. Faced with a signal $\sigma = L$ that is hurtful to his self-esteem, Self 0 chooses the recall probability λ so as to solve:

$$\max_{\lambda} \{\lambda U_T(\theta_L) + (1 - \lambda) U_C(\theta_L \mid r^*) - M(\lambda)\} \tag{15}$$

Given the convexity of $M(\lambda)$, the optimum is uniquely determined (given r^*) by the first-order condition, which involves comparing the marginal benefit from self-deception, $U_C(\theta_L \mid r^*) - U_T(\theta_L)$, with the marginal cost, $M'(\lambda)$. Finally, the Bayesian rationality of Self 1 means that he is aware of Self 0's choosing the recall strategy λ opportunistically according to (15), and uses this optimal λ in his assessment of the reliability of memories (or lack thereof).

Definition 2 A Perfect Bayesian Equilibrium (PBE) of the memory game is a pair $(\lambda^*, r^*) \in [0,1] \times [q,1]$ that solves (11) and (15), meaning that:

(i) The recall strategy of Self 0 is optimal, given Self 1's assessment of the reliability of memories.
(ii) Self 1 assesses the reliability of memories using Bayes' rule and Self 0's recall strategy.

We shall be interested in two main issues:

1 *Nature and multiplicity of equilibria*. What modes of self-esteem management are sustainable (from 'systematic denial' to 'complete self-acceptance'), depending on a person's characteristics such as his time-discounting profile or cost of memory manipulation? Can the same person, or otherwise similar people, be 'trapped' in different modes of cognition and behaviour?
2 *Welfare analysis*. Is a more active self-esteem maintenance strategy always beneficial, or can it end up being self-defeating? Would a person rather be free to manage his self-confidence and awareness as he sees fit, or prefer *a priori* to find mechanisms (friends, mates, environments, occupations, etc.) that ensure that he will always be confronted with the truth about himself, no matter how unpleasant it turns out to be?

Because PBEs are related to the solutions $r^* \in [q,1]$ to the nonlinear equation obtained by substituting (11) into the first-order condition for (15), namely

$$\psi(r,\beta) \equiv \beta\delta \int_{\beta\delta\theta_L}^{\beta\delta(r\theta_H+(1-r)\theta_L)} (\delta\theta_L - c)d\Phi(c) + M'\left(\frac{1-q/r}{1-q}\right) = 0 \tag{16}$$

we shall use a sequence of simpler cases to demonstrate the main points that emerge from our model. Note, for further reference, that $\psi(r,\beta)$ represents Self 0's (net) *marginal incentive to forget*.

Costless memory or awareness management

> 'Repression is automatic forgetting.'
> (Laughlin, *The Ego and Its Defenses,* 1979)

We first solve the model in the case where the manipulation of memory is costless, $M \equiv 0$. While it does not capture the psychological costs of repression (as opposed to the informational ones), this case already yields several key insights, and is very tractable.[22]

Proposition 2 When $M \equiv 0$, there exist $\underline{\beta}$ arid $\bar{\beta}$ in (0,1). $\underline{\beta} < \bar{\beta}$, with the following properties. For low degrees of time inconsistency, $\beta > \bar{\beta}$, the unique equilibrium involves minimum repression ($\lambda^* = 1$); for high degrees, $\beta < \underline{\beta}$, it involves maximum repression ($\lambda^* = 0$). For intermediate degrees of time inconsistency, $\beta \in [\underline{\beta}, \bar{\beta}]$, there are three equilibria, including a partially repressive one: $\lambda^* \in \{0, \Lambda(\beta), 1\}$, where $\Lambda(\beta)$ decreases from 1 to 0 as β rises from $\underline{\beta}$ to $\bar{\beta}$.

These results are illustrated in Figure 2.1. The intuition is simple, and apparent from (14). When β is high enough, overconfidence is the dominant concern, therefore adverse signals are systematically transmitted. Conversely, for low values of β the confidence-building motive dominates, so ego-threatening signals are systematically forgotten. For intermediate values both effects are relevant, allowing multiple equilibria, including one where memory is partially selective. What makes all three equilibria self-fulfilling is precisely the introspection or 'metacognition' of the Bayesian individual, who understands that his self-knowledge is subject to opportunistic distortions. The more censoring by Self 0, the more Self 1 discounts the 'no bad news to report' recollection, and therefore the lower the risk that he will be overconfident. As a result, the greater is Self 0's incentive to censor. Conversely, if Self 0 faithfully encodes all news into memory, Self 1 is more likely to be overconfident when he cannot recall any bad signals, and this incites Self 0 to be truthful.

Note that in the censoring equilibrium ($\lambda^* = 0$), none of Self 0's information is ever transmitted to Self 1: $r^* = q$. In the language of communication games, this is a 'babbling equilibrium'. The mechanism at work here is nonetheless very different from the *ex ante* suppression of information considered earlier when analysing self-handicapping, or in Carrillo and Mariotti (2000). Self 0 does not want to suppress good news, only bad news; but in doing the latter, he *cannot help* but also do the former. As we shall see later on, this

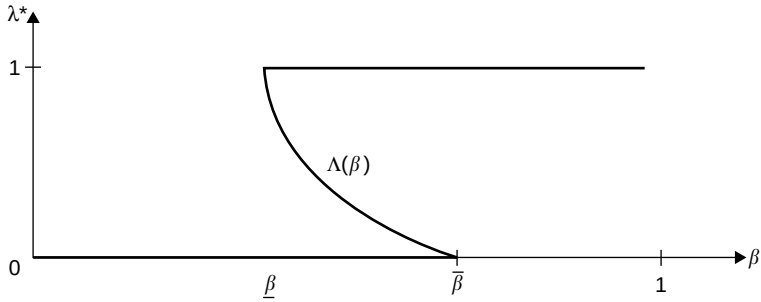


Figure 2.1 Time-consistency and equilibrium awareness (Case $M = 0$)

may end up doing him more harm than good, whereas the usual 'strategic ignorance' is only chosen when it improves *ex ante* welfare.

The last observation to be drawn from Figure 2.1 is that, as β rises from 0 to 1, there is necessarily (that is for any equilibrium selection) at least one point where λ^* has an upward discontinuity. Small differences in the psychic or material costs of memory management, repression and so on can thus imply large changes in the selectivity of memory, hence in the variability of self-confidence, and ultimately in performance.

Costly memory or awareness management

> 'To break down the renewed assaults of my memory, my imagination effectively laboured in the opposite direction.'
>
> (Marcel Proust, *Remembrance of Things Past*)

In this section we use specific functional forms to study the problem set up earlier. The memory cost function is:

$$M(\lambda) = a(1 - \ln \lambda) + b(1 - \ln(1 - \lambda)) \tag{17}$$

with $a > 0$ and $b \geq 0$. It minimized at the 'natural' recall rate $\lambda_N = a/(a+b)$, and precludes complete repression. When $b > 0$ perfect recall is also prohibitively costly, and M is U-shaped. As to the distribution of effort costs, we take it to be uniform, $\varphi(c) = 1/\bar{c}$ on $[0, \bar{c}]$, with $\bar{c} > \beta\delta\theta_H$.

With these assumptions the *incentive to forget*, namely $U_C(\theta_L \mid r^*) - U_T(\theta_L)$ in (16), is proportional to a third-degree polynomial in r, with either one or three roots in $[q, 1]$ (see the Appendix). Therefore, for any (a, b) there are again either one or three equilibria. One can go further, and obtain explicit comparative statics results, by focusing on the simpler case where *recall is costless* but *repression is costly*. The following proposition is illustrated in Figures 2.2 to 2.4.

> **Proposition 3** Let $b = 0$. A higher degree of time inconsistency or a lower cost of repression increases the scope for memory manipulation, generating partially repressive equilibria and possibly even making perfect recall unsustainable.

Formally:

1 For any given β there exist thresholds $\underline{a}$ and $\bar{a}$ with $0 \leq \underline{a} \leq \bar{a}$, and continuous functions $\lambda_1(a), \lambda_2(a)$, respectively increasing and decreasing in a, such that: (i) for $a \in (0, \underline{a})$, the unique equilibrium corresponds to $\lambda^* = \lambda_1(a)$; (ii) for $a \in (\underline{a}, \bar{a})$, there are three equilibria: $\lambda^* \in \{\lambda_1(a), \lambda_2(a), 1\}$; (iii) for $a \in (\bar{a}, +\infty)$, the unique equilibrium corresponds to $\lambda^* = 1$.

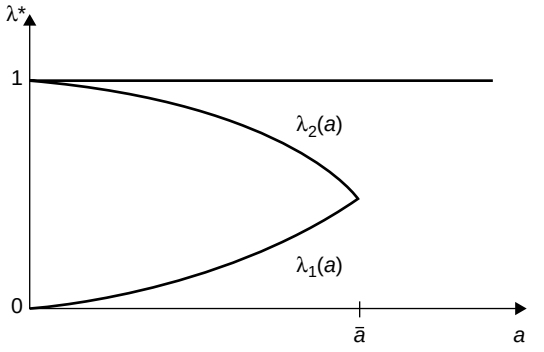


Figure 2.2 Case $\beta_2 < \beta < \beta_3$

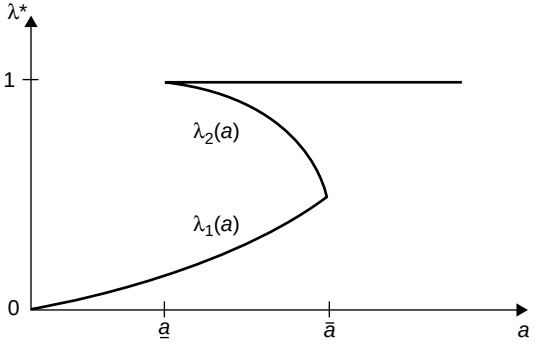


Figure 2.3 Case $\beta_1 < \beta < \beta_2$

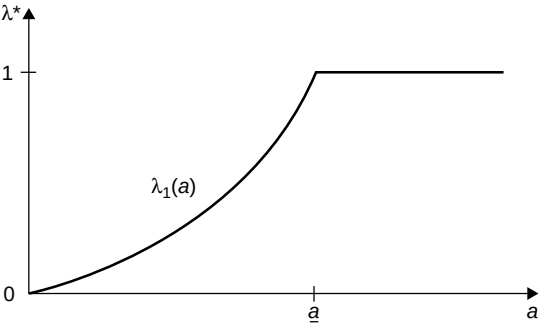


Figure 2.4 Case $\beta < \beta_1$

2 There exist critical values $\beta_1 < \beta_2 < \beta_3$ such that: (i) for $\beta \geq \beta_3$, $\bar{a} = 0$; (ii) for $\beta \in [\beta_2, \beta_3)$, $\underline{a} = 0 < \bar{a}$, as in Figure 2.2; (iii) for $\beta \in (\beta_1, \beta_2)$, $0 < \underline{a} < \bar{a}$, as in Figure 2.3; (iv) for $\beta \leq \beta_1$, $0 < \underline{a} = \bar{a}$, as in Figure 2.4.

The most representative case is that of Figure 2.3, where each of the three ranges $[0, \underline{a}]$, $[\underline{a}, \bar{a}]$, and $[\bar{a}, +\infty)$, corresponding respectively to high repression, multiplicity and truthfulness, is non-empty. The effects of a are intuitive; we just note that small changes in awareness costs can induce large changes in self-esteem and behaviour.[23] Interestingly, a lower willpower β, by shifting the equilibrium set towards lower λ's, tends to make the individual incur higher repression costs.

4 Beliefs and make-beliefs[24]

As discussed above, surveys, experiments and daily observation consistently suggest that most people overestimate their past achievements, abilities and other desirable traits, both in absolute terms and relative to others (see for example Weinstein, 1980; Taylor and Brown, 1988). Well-educated, reflective individuals seem to be no exception: as Gilovich (1991) relates, 'a survey of college professors found that 94% thought they were better than their average colleague'. These findings are often put forward as evidence of pervasive irrationality in human inference.[25] It turns out, however, that rational self-deception by Bayesian agents can help account for most people holding biased, self-serving beliefs, which in turn have aggregate effects.

Optimistic and pessimistic biases

Continuing to work with our awareness-management model, let us compare the cross-sectional distributions of true and self-perceived abilities.[26] In a large population, a proportion $1 - q$ of individuals are of low ability θ_L, having received a negative signal, $\sigma = L$. The remaining q, having received $\sigma = \emptyset$, have high ability θ_H. Average ability is $q\theta_H + (1-q)\theta_L = \theta(q)$; we assume that $q < 1/2$, so that median ability is θ_L. Consider now the distribution of self-evaluations. Suppose for simplicity that, when faced with ego-threatening information ($\sigma = L$), everyone uses the same censoring probability $\lambda^* \in (0, 1)$.[27] As before, let r^* denote the corresponding reliability of memory. given by (11). Thus, when individuals make decisions at date 1,

> a fraction $(1-q)(1-\lambda^*)$ *overestimate* their ability by $\theta(r^*) - \theta_L = r^*(\theta_H - \theta_L)$;
> a fraction q *underestimate* it by $\theta_H - \theta(r^*) = (1 - r^*)(\theta_H - \theta_L)$.

If the costs of repression or forgetting are low enough (for example a small a in Figure 2.4), one can easily have $(1 - q)(1 - \lambda^*) > 1/2$, and even $(1 - q)(1 - \lambda^*) \gtrapprox 1$. Thus most people believe themselves to be *more able than they actually are, more able than average, and more able than the majority*

of individuals.[28] Adding those who had truly received good news ($\sigma = \emptyset$), the fraction of the population who think they are better than average is even larger, namely $1 - \lambda^*(1-q)$. The remaining minority think, correctly, that they are worse than average; as a result they have low motivation and are unlikely to undertake challenging tasks. They fit the experimental findings of depressed people as 'sadder but wiser' realists, compared with their non-depressed counterparts who are much more likely to exhibit self-serving delusions (Alloy and Abrahamson, 1979).

As seen above, Bayes' law does not constrain the skewness in the distribution of biases:[29] it only requires that the *average* bias across the $(1-q)(1-\lambda^*)$ optimists and the q pessimists be zero: indeed, $(1-q)(1-\lambda^*)r^* - q(1-r^*) = 0$ by (11). In other words, Bayesian rationality only imposes a trade-off between the relative proportions of overconfident versus underconfident agents in the population, and their respective degrees of over- or under-confidence. Note, however, that a zero average bias in no way precludes self-esteem maintenance strategies from having aggregate economic effects. Clearly, in our model they do affect aggregate effort, output and welfare, as none of these is a linear function of perceived ability.

To Bayes or not to Bayes?

Having shown that even rational agents can deceive themselves most of the time (albeit not all the time) we nonetheless recognize that it may be more realistic to view people as *imperfect Bayesians* who do not fully internalize the fact that their recollections may be self-serving. At the other extreme, taking beliefs as completely naive would be even more implausible. As argued earlier (p. 32; also recall Nietzsche and Darwin), if a person consistently destroys, represses or manages not to think about negative news, he will likely become aware that he has this systematic tendency, and realize that the absence of adverse evidence or recollections should not be taken at face value. This introspection is the fundamental trait of the human mind which the Bayesian assumption captures in our model. Without it, self-delusion would be very easy and, when practiced, always optimal (*ex ante*). With even *some* of this metacognition, self-deception becomes a much more subtle and complex endeavour.

In Bénabou and Tirole (1999) we relax Bayesian rationality and allow the agent at date 1 to remain unaware not just of *what* he may have forgotten, but also of the fact *that* he forgets. Self 1's assessment of the reliability of a recollection $\hat{\sigma} = \emptyset$ is thus modified to:

$$r_\pi(\lambda) \equiv \Pr[\sigma = \emptyset \,|\, \hat{\sigma} = \emptyset; \lambda] = \frac{q}{q + \pi(1-q)(1-\lambda)} \tag{18}$$

where λ is the actual recall strategy and $\pi \in [0,1]$ parameterizes cognitive sophistication, ranging from complete naivete ($r_0(\lambda) \equiv 1$) to full rationality

($r_1(\lambda) \equiv r^*(\lambda)$). As long as π is above a critical threshold, meaning that the individual's self-conception is not too unresponsive to his actual pattern of behaviour, all the Bayesian model's results on multiplicity and welfare rankings of personal equilibria go through. Thus, for the purpose of understanding self-deception and overoptimism, the explanatory power gained by departing from rational inference is rather limited (perception biases need no longer sum to zero), whereas much can be lost if the departure is too drastic: without sufficient introspection, one cannot account for 'self-traps' or self-doubt.

5 Welfare analysis of self-deception

> 'The art of being wise is the art, of knowing what to overlook.'
> (William James, *Principles of Psychology*, 1890)

> 'There is nothing worse than self-deception – when the deceiver is at home and always with you.' (Plato, quoted by Mele, 1997)

Is a person ultimately better off in an equilibrium with a strategy of active self-esteem maintenance and 'positive thinking' ($\lambda^* < 1$), or when he always faces the truth? Like Plato and William James, psychologists are divided between these two conflicting views of self-deception. On one side are those who endorse and actively promote the self-efficacy/self-esteem movement (for example Bandura, 1977, and Seligman, 1990), pointing to studies which tend to show that a moderate dose of 'positive illusions' has significant affective and functional benefits.[30] On the other side are skeptics and outright critics (for example Baumeister, 1998, and Swann, 1996), who see instead a lack of convincing evidence, and point to the dangers of overconfidence as well as the loss of standards that results when negative feedback is systematically withheld in the name of self-esteem preservation. Our model will provide insights into the reasons for this ambiguity.

Consider an equilibrium with recall probability $\lambda^* \leq 1$ and associated credibility r^* (via (11)). With probability $1 - q$, Self 0 receives bad news, which he then forgets with probability $1 - \lambda^*$; the resulting expected payoff is $\lambda^* U_T(\theta_L) + (1 - \lambda^*) U_C(\theta_L \,|\, r^*) - M(\lambda^*)$. With probability q the news are good, which means that no adverse signal is received. The problem is that the credibility of a 'no bad news' memory in the eyes of Self 1 may be quite low, so that he will not exert much effort even when it is actually optimal to do so. Indeed, the payoff to Self 0 following genuinely 'good news' is only:

$$U_T(\theta_H \,|\, r^*) = \beta\delta \int_0^{\beta\delta\theta(r^*)} (\delta\theta_H - c) d\Phi(c) \tag{19}$$

which is clearly less than $U_T(\theta_H \mid 1)$ whenever cost-realizations between $\theta(r^*)$ and $\theta(1)$ have positive probability. In that case there is a loss from *self-distrust* or *self-doubt*, compared with a situation where Self 0 always truthfully records all events into memory. Like a ruler whose entourage dares not bring him bad news, or a child whose parents praise him indiscriminately, an individual with some understanding of the self-serving tendency in his attention or memory can never be sure that he really 'did great', even in instances where this was actually true.

Averaging over good and bad news, the agent's *ex ante* welfare in equilibrium equals:

$$\mathcal{W}(\lambda^*, r^*) \equiv qU_T(\theta_H \mid r^*) + (1-q)[\lambda^* U_T(\theta_L) + (1-\lambda^*)U_C(\theta_L \mid r^*) - M(\lambda^*)] \tag{20}$$

Let us now assume that truth (perfect recall) is also an equilibrium strategy, with cost $M(1)$; as we shall see, a very similar analysis applies if $\lambda^* = 1$ is achieved by using some *a priori* commitment mechanism (chosen before σ is observed). Denoting the difference in welfare with this benchmark case as $\Delta W(\lambda^*, r^*) \equiv \mathcal{W}(\lambda^*, r^*) - \mathcal{W}(1,1)$, we have:

$$\begin{aligned}\Delta W(\lambda^*, r^*) = {} & (1-q)[(1-\lambda^*)(U_C(\theta_L \mid r^*) - U_T(\theta_L)) - M(\lambda^*) + M(1)] \\ & - q[U_T(\theta_H \mid 1) - U_T(\theta_H \mid r^*)]\end{aligned}$$

or, equivalently:

$$\begin{aligned}\Delta W(\lambda^*, r^*) = {} & (1-q)\left((1-\lambda^*)\int_{\beta\delta\theta_L}^{\beta\delta\theta(r^*)} (\delta\theta_L - c)d\Phi(c) - M(\lambda^*) + M(1)\right) \\ & - q\int_{\beta\delta\theta(r^*)}^{\beta\delta\theta_H} (\delta\theta_H - c)d\Phi(c)\end{aligned} \tag{21}$$

The first expression describes the net *gain from forgetting bad news*, the second one the *loss from disbelieving good news*. While the individual is better motivated or even over-motivated following a negative signal about his ability, he may actually be undermotivated following a good signal. A few general results can be immediately observed.

First, if memory manipulation is costless ($M \equiv 0$), then a partial recall (mixed strategy) equilibrium can never be better than perfect recall. Indeed, in such an equilibrium the gain from hiding bad news is zero ($U_C(\theta_L \mid r^*) = U_T(\theta_L)$), because the self-enhancement and overconfidence effects just cancel out. The cost from self-distrust, on the other hand, is always present.

When repression is costly this reasoning no longer applies, as the term in large brackets in (21) becomes $M(1) - M(\lambda^*) - (1-\lambda^*)M'(\lambda^*) > 0$, by the convexity of M. Similarly, when systematic denial ($\lambda^* = 0$) is an equilibrium,

it generates a positive 'surplus' in the event of bad news: $U_C(\theta_L \mid q) - U_T(\theta_L) > M'(0) \geq 0$. How does this gain compare with the loss from self-distrust in the good-news state? As seen from (21), the key intuition involves the likelihood of cost realizations sufficiently high to discourage effort in the absence of adverse recollections ($\hat{\sigma} = \emptyset$). When such events are infrequent, the self-distrust effect is small or even absent and, on average, self-deception pays off. When they are relatively common, the reverse is true.

Proposition 4 Let $M \equiv 0$. If the cost density $\varphi(c)$ decreases fast enough,

$$-\frac{\partial \ln \varphi(c)}{\partial \ln c} > \frac{2-\beta}{1-\beta} \quad \text{for all } c \in [0, \bar{c}],$$

then *ex ante* welfare is higher if all bad news is censored from memory than if it is always recalled. If the inequality is reversed, so is the welfare ranking. For a given cost distribution φ, self-deception is thus more likely to be beneficial for a less time-consistent individual.

Note that even the second result was far from obvious *a priori*, since both the gain and the loss in (21) decrease with β: in equilibrium, memory manipulation tends to alleviate procrastination when $\sigma = L$, but worsen it when $\sigma = \emptyset$.[31] To summarize, we have shown that:

1 When the tasks one faces are very difficult and one's willpower is not that strong, a strategy of active self-esteem maintenance, 'looking on the bright side', avoiding 'negative' thoughts and people, etc., as advocated in numerous 'self-help' books, can indeed pay off.
2 When the typical task is likely to be only moderately challenging, and time-inconsistency is relatively mild, one can only lose by playing such games with oneself, and it would be better to always 'be honest with yourself' and 'accept who you are'.

It is important to note that, in the second case, the individual *may still play* such denial games, even though self-honesty would be better. First, he could be trapped in an inferior equilibrium. Second, motivated cognition may be the only equilibrium, yet still result in lower welfare than if the individual could commit to never try and fool himself.[32] A couple of examples will help make these results more concrete:

(a) With a uniform density on $[0, \bar{c}]$ ($\bar{c} > \beta\delta\theta_H$), self-deception is *always harmful* compared with truth-telling. This applies whether both $\lambda^* = 0$ and $\lambda^* = 1$ are in the equilibrium set, or only one of them (see Proposition 2 and Figure 2.1). It also applies, *a fortiori*, when repression is costly.

(b) Conversely, self-deception is *always beneficial* when $\varphi(c) = \gamma c^{-n}$ on $[\underline{c}, +\infty]$, with $0 < \underline{c} < \beta\delta\theta_L, \gamma$ chosen so that the density sums to one, and $n > (2-\beta)/(1-\beta)$. In this case it can also be shown that $\lambda^* = 0$ is the only equilibrium if $M \neq 0$.
(c) Finally, we provide in the Appendix a simple example where both $\lambda^* = 0$ and $\lambda^* = 1$ coexist as equilibria and where *either one* – depending on parameter values – may lead to higher *ex ante* welfare.

We have thus far interpreted the 'always face the truth' strategy as an equilibrium, sustainable alongside with λ^*. Alternatively, it could result from some initial commitment of the type discussed earlier (chosen before σ is observed), which amounts to *making oneself face steeper costs of self-deception* (increasing $M(\lambda)$ for $\lambda < \lambda_N$).[33] This reinterpretation requires minor modifications to (21), but the main conclusions remain unaltered.[34]

The potential multiplicity of equilibria in our model raises the issue of coordination among the individual's temporal selves. Observe that Self 1 always values information about the productivity of his own efforts, and therefore always ranks equilibria (or commitment outcomes) in order of increasing λ's. When Self 0 also prefers the $\lambda^* = 1$ solution, it is plausible (we are agnostic on this point) that the individual will find ways to coordinate on this Pareto-superior outcome. When some repression (any $\lambda^* < 1$) is *ex ante* valuable, however, there is no longer any clearly natural selection rule. In either case, our main welfare conclusions remain unchanged even if one assumes that Self 0 always manages to select his preferred equilibrium. First, for some range of β or a, $\lambda^* = 1$ ceases to be an equilibrium even though its still maximizes *ex ante* welfare. Thus, once again, the individual is trapped in a harmful pattern of systematic denial. Conversely, for relatively high values of a the only equilibrium may be $\lambda^* = 1$ (more generally, a high λ^*), even though the individual would, *ex ante*, be better off if he could manage to repress bad news more easily.

Interestingly, our multiplicity and welfare results provide a role for parents, friends, therapists and other benevolent outside parties to help an individual escape the 'self-traps' (Swann, 1996) in which he might be stuck: a depressive state of low self-esteem, chronic blindness to his own failings and so on. They can make him aware that a better personal equilibrium is feasible, and teach him how to coordinate on it by following certain simple cognitive rules. They may also offer a form of informational commitment, serving as the repositories of facts and feelings which the individual realizes that he has an incentive to forget ('let's talk about that incident with your mother again'). More generally, they allow him to alter the 'awareness/repression' technology $M(\lambda)$ (and hence the set of feasible equilibria), whether through their own feedback and questioning, or by teaching him certain cue-management techniques. Indeed, much of modern cognitive therapy aims at changing people's self-image through selective recollection and rehearsal of events,

self-serving attributions about success and adversity, or conversely helping them 'see through' harmful self-delusions.

6 Variants and extensions

Defensive pessimism

While people are most often concerned with enhancing and protecting their self-esteem, there are also many instances where they seek to minimize their achievements, or convince themselves that the task at hand will be difficult rather than easy. A student preparing for exams may thus discount his previous good grades as attributable to luck or lack of difficulty. A young researcher may understate the value of his prior achievements, compared with what will be required to obtain tenure. A dieting person who lost a moderate amount of weight may decide that he 'looks fatter than ever', no matter what others or the scale may say.

Such behaviour, termed 'defensive pessimism' by psychologists, can be captured with a very simple variant of our model. The above are situations where the underlying motive for information-manipulation is still the same, namely to alleviate the shirking incentives of future selves; the only difference is that ability is now a *substitute* rather than a complement to effort in generating future payoffs. This gives the agent an incentive to discount, ignore and otherwise repress signals of *high* ability, as these would increase the temptation to 'coast' or 'slack off'.

Substitutability may arise directly in the performance 'production function' which, instead of the multiplicative form $\pi(e,\theta) = \theta e$ that we assumed, could be of the form $\pi(e,\theta)$ with $\pi_{e\theta} < 0$. More interestingly, it will typically occur when the *reward* for performance is of a 'pass–fail' nature: graduating from school, making a sale, being hired or fired (tenure, partnership), proposing marriage and so on. To see this, let performance remain multiplicative in ability and effort: $\pi(\theta, e, \varepsilon) = \varepsilon\theta e$, where ε is a random shock with cumulative distribution $H(\varepsilon)$. The payoff V, however, is now conditional on performance exceeding a cutoff level $\bar{\pi}$. Self 1's utility function is thus

$$\beta\delta V \Pr[\varepsilon\theta e \geq \bar{\pi}] - ce = \beta\delta V(1 - H(\bar{\pi}/\theta e)) - ce \tag{22}$$

It is easily verified that if the density $h = H'$ is such that $xh'(x)/h(x) > -1$ on the relevant range of $x \equiv \bar{\pi}/\theta e$, the optimal effort is decreasing in θ. Note that these results yield *testable* predictions: by comparing subjects' confidence-maintenance behaviour across experiments (or careers) where payoffs are complements and substitutes, one should be able to distinguish between the motivation-based theory of self-confidence and the hedonic or signalling alternatives.

An even simpler form of defensive pessimism arises in situations where the action subject to time inconsistency is such that *the benefits precede the costs*. One can think of the trade-off between immediate pleasure of smoking, drinking, spending freely and so on, and the long term, large but uncertain costs of such behaviours. Suppose, for instance, that, at date 1 the decision is to consume or not consume. The first option yields utility b, but with probability ω entails a cost C at date 2; the second option yields zero at both dates. Clearly, if we define 'effort' $e \equiv 1 - x$ as abstinence from consumption, $c = b$ as its (opportunity) cost and ωC as its expected long-term payoff, we see that this problem fits exactly with our model. Thus, to counteract his tendency towards short-term gratification, the agent will try and maintain beliefs that ω and C are high.[35] Note that these variables are complements to $e = 1 - x$ in his utility function. If we had framed the problem in terms of uncertainty over the probability of being immune to the health risks (say) from tobacco or alcohol, this probability $1 - \omega$ would be a substitute with e, so the agent would like to understate it to his future selves. Whether costs precede or follow benefits thus simply amounts to a relabelling of variables. The only thing that matters for the direction in which the agent would like to manipulate his beliefs concerning a variable is its cross-derivative with the decision variable that is being set inefficiently low due to time inconsistency.

Self-esteem as a consumption good

We have until now emphasized the value of self-confidence for personal motivation. This approach provides an explanation of both *why* and *how much* people care about their self-image: its value arises endogenously from fundamental preferences, technological constraints and the structure of incentives. As explained earlier, the motivation theory also readily extends to social interactions.

This functional view of self-esteem, while pervasive in psychology, is by no means the only one (see for example Baumeister, 1998). As discussed earlier, a common and complementary view involves purely affective concerns: people just *like* to think of themselves as good, able, generous and attractive, and conversely find it painful to contemplate their failures and shortcomings. Formally, self-image is simply posited to be an argument of the utility function. This potentially allows people to care about a broader set of self-attributes than a purely motivation-based theory: they may for instance want to perceive themselves as honest and compassionate individuals, good citizens, faithful spouses and so on or, on the contrary, pride themselves on being ruthless businessmen, ultrarational economists, irresistible seducers, for example. There is somewhat of an embarrassment of riches here, with few constraints on what arguments should enter the utility function, and with what sign.

Let us therefore focus, as before, on the trait of 'general ability', which presumably everyone views as a good. This is also the type of attribute from which agents are assumed to derive utility in Weinberg (1999) and Köszegi (1999), as well as in some interpretations which Akerlof and Dickens (1982) offered for their model of dissonance reduction. The trade-off between the costs and benefits of information can then be modelled by positing preferences of the form:

$$\mathsf{E}[\max\{\hat{\theta}V - c, 0\} + u(\hat{\theta})] \tag{23}$$

where $\hat{\theta}$ denotes the individual's self-perceived ability (expected probability of success) at the time of the effort decision.[36] The first term always generates a demand for accurate information, to improve decision-making. Suppose for now that the hedonic valuation $u(\hat{\theta})$ is increasing and concave; these properties respectively imply a positive demand for self-esteem, and risk-aversion with respect to self-relevant signals. The individual may then, once again, avoid free information or engage in self-handicapping.[37] Similarly, all our results based on memory management on the supply side carry over to this case. Thus, 'positive thinking' and similar self-deception strategies may be pursued even though they are ultimately detrimental (recall the quotation from Plato), while conversely personal rules not to tamper with the encoding and recall of information, such as Darwin's, can be valuable. The basic insight is, again, one of *externalities across information states*: having only good news is not such a great boost to self-esteem once the agent realizes that he would have had reasons to censor any bad news that might have been received.

Unfortunately, psychology provides little guidance on what the appropriate shape of the hedonic preference function should be (by contrast, there is ample evidence of people's general bias towards short-term rewards, tendency to procrastinate and so on). It is thus equally likely that $u(\hat{\theta})$ is convex, at least over some range; in such cases the individual will be an avid information-seeker, choosing tasks that are excessively hard or risky but very informative, as a way of gambling for (self) resurrection. Even monotonicity may not be taken for granted, since psychologists have documented both optimism and defensive pessimism. The latter, whether originating from motivation concerns or hedonic ones (lowering one's expectations of performance because surprise sharpens both the sweetness of success and the bitterness of defeat), requires that $u(\hat{\theta})$ be sometimes decreasing.

7 Conclusion

Building on a number of themes from cognitive and social psychology, we have proposed in this chapter a general economic model of why people value their self-image, and of how they seek to enhance or preserve it through a variety of seemingly irrational behaviours – from handicapping their

own performance to practicing self-deception through selective memory or awareness management.

This general framework can be enriched in many directions. On the motivation side, we noted earlier that anyone with a vested interest in an individual's success (or failure) has incentives to manipulate the latter's self-perception. Thus, in principal–agent relationships or bargaining situations, the management of self-confidence will matter even when everyone is fully time-consistent. These issues are explored in Bénabou and Tirole (2003), where we examine the provision of incentives by informed principals (parents, teachers, managers) in educational and workplace environments. Because offering rewards for performance may signal low trust in the abilities of the agent (child, student, worker) or in his suitability to the task, such extrinsic motivators may have only a limited impact on his current performance, and undermine his intrinsic motivation for similar tasks in the future – as stressed by psychologists.

Another interesting direction is to further explore the rich set of behavioural implications that arise from the interaction of imperfect willpower and imperfect memory. Thus, in Bénabou and Tirole (2004) we develop a model of self-reputation over one's willpower that can account for the 'personal rules' (diets, exercise regimens, resolutions, moral or religious precepts and so on) through which people attempt to achieve self-discipline. The sustainability of rule-based behaviour is shown to depend on the effectiveness of the individual's *self-monitoring* (recalling past lapses and their proper interpretation), which may be subject to opportunistic distortions of memory or inference of the type studied here. The model also helps explain why people may sometimes adopt excessively 'legalistic' rules that result in compulsive behaviour such as miserliness, workaholism or anorexia.

Appendix

Proof of Proposition 2

For all r and β in [0,1], let us define:

$$\chi(r,\beta) \equiv \int_{\beta\delta\theta_L}^{\beta\delta\theta(r)} (\delta\theta_L - c)d\Phi(c) \tag{A.1}$$

which, up to a factor of $\beta\delta$, measures the incentive to forget bad news, $U_C(\theta_L \mid r^*) - U_T(\theta_L)$.

Lemma 1 For all $r \in [0,1]$, there exists a unique $B(r) \in [0,1]$ such that $\chi(r, B(r)) = 0$ and:

(i) $\chi(r,\beta) > 0$ for all $\beta < B(r)$, while $\chi(r,\beta) < 0$ for all $\beta > B(r)$;
(ii) $B(r) > \theta_L/\theta(r)$, and $B(r)$ is strictly decreasing in r.

Proof For any given r, it is clear from (A.1) that $\chi(r,\beta) > 0$ for $\beta \in [0, \theta_L/\theta(r)]$, while $\chi(r,1) < 0$. Moreover, for all $\beta > \theta_L/\theta(r)$, we have:

$$\frac{\partial \chi(r,\beta)}{\partial \beta} = \delta^2\theta(r)[\theta_L - \beta\theta(r)]\varphi(\beta\delta\theta(r)) - \delta^2\theta_L[\theta_L - \beta\theta_L]\varphi(\beta\delta\theta_L) < 0 \tag{A.2}$$

This establishes the existence and uniqueness of the root $B(r) \in [\theta_L/\theta(r), 1]$. Moreover,

$$\frac{\partial \chi(r,\beta)}{\partial r} = \beta\delta^2(\theta_H - \theta_L)[\theta_L - \beta\theta(r)]\varphi(\beta\delta\theta(r)) \tag{A.3}$$

so $\partial\chi(r, B(r))/\partial r < 0$ since $B(r)\theta(r) > \theta_L$. Therefore, by the implicit function theorem, $B'(r) < 0$ for all r. ‖

To conclude the proof of Proposition 2, consider the following cases:

(a) For $\beta \geq B(q)$ we have, for all $r \in [q,1]$, $\beta > B(r)$ and therefore $\chi(r,\beta) < 0$. Memorizing bad news is thus the optimal strategy, which establishes claim (i) of the Proposition.

(b) For $\beta \leq B(1)$ we have, for all $r \in [q,1]$, $\beta < B(r)$ and therefore $\chi(r,\beta) > 0$. Forgetting bad news is thus the optimal strategy, which establishes claim (ii).

(c) For $\beta \in (B(1), B(q))$ there exists by the lemma a unique inverse function $R(\beta) \equiv B^{-1}(\beta)$, such that $\chi(R(\beta),\beta) = 0$. Moreover, R is decreasing and for any $r \in (q,1)$, $\chi(r,\beta)$ has the sign of $R(\beta) - r$. Therefore, the only equilibrium with $r < R(\beta)$ is $r = q$ (or $\lambda = 0$), and the only equilibrium with $r > R(\beta)$ is $r = 1$, (or $\lambda = 1$). Finally, $r = R(\beta)$ is also an equilibrium, which corresponds to $\Lambda(\beta) = (1 - q/R(\beta))(1 - q)$. Defining $\underline{\beta} \equiv B(1)$ and $\bar{\beta} \equiv B(q)$ concludes the proof.

Proof of Proposition 3

We shall solve for equilibria in terms of the reliability of memory, r^*; the recall strategy λ^* is then obtained by inverting (11). With the assumptions of the proposition, the incentive to forget, given by (16), equals:

$$\psi(r,\beta) = r(\Delta\theta)\left(\frac{\beta^2\delta^3}{\bar{c}}\right)\left((1-\beta)\theta_L - \frac{\beta r}{2}(\Delta\theta)\right)r$$
$$- ar\left(\frac{1-q}{r-q}\right) + br\left(\frac{1-q}{q(1-r)}\right) \tag{A.4}$$

where $\Delta\theta \equiv \theta_H - \theta_L$. Defining, for all $\beta \in [0,1]$:

$$R(\beta) \equiv \left(\frac{1-\beta}{\beta}\right)\frac{2\theta_L}{\Delta\theta} \tag{A.5}$$

$$\Omega(\beta) \equiv (\Delta\theta)^2\left(\frac{\beta^3\delta^3}{2\bar{c}}\right)\left(\frac{q}{1-q}\right) \tag{A.6}$$

it is clear that $\psi(r,\beta) \geq 0$ if and only if:

$$P(r,\beta) \equiv \Omega(\beta)(R(\beta) - r)(r - q) \geq aq - b\left(\frac{r-q}{1-r}\right) \tag{A.7}$$

Multiplying by $(1-r)$ shows that the sign of $\psi(r,\beta)$ is that of a third-degree polynomial in r, with $\lim_{r\to q}\psi(r,\beta) = -\infty$ since $a > 0$, and $\lim_{r\to 1}\psi(r,\beta) = +\infty$ when $b > 0$. Thus,

for a given β there are either one or three solutions to $\psi(r,\beta)=0$ in $[q,1]$, that is one or three equilibria.

Let us now specialize (A.7) to the case where remembering is costless but forgetting or repressing is costly, $b=0$. Solving $\psi(r,\beta)=0$ then reduces to looking for the intersections of the *quadratic* polynomial $P(r,\beta)$ with the horizontal line aq. We shall denote $\bar{a}\equiv q^{-1}\max\{\max_{r\in[q,1]}P(r,\beta),0\}$ and $\underline{a}\equiv q^{-1}\max\{P(1,\beta),0\}\leq\bar{a}$. There are several cases to consider:

1 For $R(\beta)<q$, or equivalently $\beta>R^{-1}(q)\equiv\beta_3$, it is clear that $P(r,\beta)<0$ on $[q,1]$, therefore the only equilibrium is $r=1$. Moreover, $\underline{a}=\bar{a}=0$.
2 For $q<R(\beta)\equiv\beta_3$ the polynomial $P(r,\beta)$ is positive on $r\in[q,R(\beta)]$, implying $\bar{a}>0$, and negative outside.
   (a) If $a>\bar{a}$, then $P(r,\beta)<0$ on $[q,R(\beta)]$, so the only equilibrium is again $r=1$.
   (b) If $a\leq\bar{a}$, the equation $P(r,\beta)=aq$ has two roots $r_1(a)$ and $r_2(a)$, both in the interval $[q,R(\beta)]$, with $r_1(a)\leq r_2(a)$, r_1 decreasing and r_2 increasing. With these one can associate two functions, $\lambda_1(a)$ and $\lambda_2(a)$, by inverting (11).

Let us now distinguish the following subcases:

(i) For $q<R(\beta)<1$, or equivalently $\beta_2\equiv R^{-1}(1)<\beta<R^{-1}(q)=\beta_3$, both $r_1(a)$ and $r_2(a)$ are in $(q,R(\beta))$ and represent equilibria. On $[(q,r_1(a))$ and $(r_2(a),1)]$ we have $P(r,\beta)<aq$, hence $\psi(r,\beta)<0$. This means that the third (and only other) equilibrium is $r=1$. Furthermore, $\underline{a}=0$.
(ii) For $1<R(\beta)<2-q$, or equivalently $\beta_1\equiv R^{-1}(2-q)<\beta<\beta_2=R^{-1}(1)$, the polynomial $P(r,\beta)$ reaches its maximum at $(q+R(\beta))/2<1$. Thus $P(r,\beta)$ is positive and hill-shaped on $[q,1]$, and $\underline{a}=P(1,\beta)>0$. This implies that for $\underline{a}<a<\bar{a}$ we have $q<r_1(a)<r_2(a)<1$, while for $a<\underline{a}$ we have $q<r_1(a)<1<r_2(a)$. In the first case the equilibria are $r\in\{r_1(a),r_2(a),1\}$, as in case (i) above. In the latter situation the only equilibrium is $r=r_1(a)$.
(iii) For $2-q<R(\beta)$, or equivalently $\beta<\beta_1\equiv R^{-1}(2-q)$, the polynomial $P(r,\beta)$ is strictly increasing on $[q,1]$, so the only equilibrium is $r=r_1(a)$ whenever $a<\underline{a}=P(1,\beta)=\bar{a}$. It is $r=1$ whenever $a\geq\underline{a}$.

Proof of Proposition 4

Setting $\lambda^*=0, r^*=q$ and $M\equiv 0$ in (21) yields:

$$\begin{aligned}\Delta W(0,q)&=(1-q)\int_{\beta\delta\theta_L}^{\beta\delta\theta(q)}(\delta\theta_L-c)d\Phi(c)-q\int_{\beta\delta\theta(q)}^{\beta\delta\theta_H}(\delta\theta_H-c)d\Phi(c)\\&=q\int_0^{\beta\delta\theta(q)}(\delta\theta_H-c)d\Phi(c)+(1-q)\int_0^{\beta\delta\theta(q)}(\delta\theta_L-c)d\Phi(c)\\&\quad-q\int_0^{\beta\delta\theta_H}(\delta\theta_H-c)d\Phi(c)-(1-q)\int_0^{\beta\delta\theta_L}(\delta\theta_L-c)d\Phi(c)\\&=\int_0^{\beta\delta\theta(q)}[\delta(q\theta_H+(1-q)\theta_L)-c]d\Phi(c)\\&\quad-q\int_0^{\beta\delta\theta_H}(\delta\theta_H-c)d\Phi(c)-(1-q)\int_0^{\beta\delta\theta_L}(\delta\theta_L-c)d\Phi(c)\end{aligned}$$

Defining the function $\Gamma(Z,\beta)\equiv\int_0^Z(Z-\beta c)d\Phi(c)$, we can then write

$$\Delta W(0,q)=\beta^{-1}[\Gamma(\beta\delta(q\theta_H+(1-q)\theta_L),\beta)-q\Gamma(\beta\delta\theta_H,\beta)-(1-q)\Gamma(\beta\delta\theta_L,\beta)] \quad (A.8)$$

Clearly, $\Delta W(0,q) > 0$ when Γ is concave in Z, and $\Delta W(0,q) < 0$ when it is convex. Indeed, $\beta\delta\Delta W(0,q)$ is (minus) the *ex ante* value of information, that is of always knowing the true $E[\theta \mid \sigma]$ rather than have only the uninformed prior or posterior $\theta(q)$. The proposition immediately follows from the fact that $\partial^2\Gamma(Z,\beta)/\partial Z^2 = (2-\beta)\varphi(Z) + (1-\beta)Z\varphi'(Z)$.

Welfare rankings of multiple equilibria

We construct here a simple example where $\lambda^* = 0$ and $\lambda^* = 1$ coexist as equilibria, and where *either one* can lead to higher *ex ante* welfare.

First, let $\theta_L < \theta_H$ and $q \in (0,1)$, so that $\theta_L < \theta(q) = q\theta_H + (1-q)\theta_L < \theta_H$. For $\beta < 1$ but not too small we have $\beta\theta_L < \theta_L < \beta\theta(q) < \theta(q) < \beta\theta_H < \theta_H$. Next, let the date-1 cost take two values: $c \in \{\underline{c}, \bar{c}\}$, with $\underline{c}/\delta \in (\beta\theta_L, \theta_L)$, $\bar{c}/\delta \in (\theta(q), \beta\theta_H)$ and $\pi \equiv \Pr[c = \underline{c}] \in (0,1)$. The $\lambda^* = 0$ strategy is then always an equilibrium, since $\psi(q,\beta)/\beta\delta = \pi(\delta\theta_L - \underline{c}) > 0$. As to $\lambda^* = 1$, it is also an equilibrium whenever $\psi(1,\beta)/\beta\delta = \pi(\delta\theta_L - \underline{c}) - (1-\pi)(\bar{c} - \delta\theta_L) < 0$, or

$$\frac{\pi}{1-\pi} < \frac{\bar{c} - \delta\theta_L}{\delta\theta_L - \underline{c}} \equiv \rho \tag{A.9}$$

With this condition both $\lambda^* = 0$ and $\lambda^* = 1$ are equilibria (with a mixed-strategy one in between, which is always dominated by $\lambda^* = 1$ since $M \equiv 0$), and $\lambda^* = 0$ yields higher welfare when:

$$\Delta W(0,\beta) = (1-q)\pi(\delta\theta_L - \underline{c}) - q(1-\pi)(\delta\theta_H - \bar{c}) > 0$$

or

$$\frac{\pi}{1-\pi} > \left(\frac{q}{1-q}\right)\left(\frac{\delta\theta_H - \bar{c}}{\delta\theta_L - \underline{c}}\right) \equiv \rho' \tag{A.10}$$

Since $\theta(q) < \delta\bar{c}$, it is easily verified that $\rho' < \rho$. Thus for $\pi/(1-\pi) \in (\rho', \rho)$, the $\lambda = 0$ equilibrium is *ex ante* superior to the one with $\lambda = 1$. For $\pi/(1-\pi) < \rho'$ the reverse is true.

Notes

1 The psychology literature generally views introspection as quite inaccurate (Nisbett and Wilson, 1977), and stresses that learning about oneself is an ongoing process. Furthermore, the self is constantly changing (e.g., Rhodenwalt, 1986): personal characteristics evolve with age, the goals pursued shift over one's career and life-cycle (often as the result of interactions with others), and the personal or economic environment in which these objectives are rewarded is typically variable.

2 Thus, Gilbert and Cooper (1985) note that 'the classic attributional model of the causes of behaviour... [is described by] the well-known conceptual equation: $(E \times A) \pm TD = B$, in which effort times ability, plus or minus task difficulty equals the behavioral outcome.' Additional references are given in section 2. Note, however, that there are also instances where ability and effort are substitutes. As discussed below, we consider this case as well.

3 Thus *attribution theory* (Heider, 1958) emphasizes the distinction between temporary (situational) and enduring (dispositional) characteristics. In economics parlance, the individual filters out noise in order to extract information from past events. In the *social comparison process* (Festinger, 1954), individuals assess their ability by comparing their performance with that of people facing similar

conditions (familial, cultural, educational, etc.,). In other words, they use 'relative performance evaluation', or 'benchmarking', for self-evaluation. A good performance by others in one's reference group is thus generally detrimental to self-esteem, and conversely some comfort is derived when others experience adversity (*Schadenfreude*). Relatively sophisticated updating also applies to the interpretation of praise and criticism: a person takes into account not only what others say (or do), but also their possible intentions.

4 Why they would want to do so in a social context is obvious. The interesting question is why they may bias their own inference process.

5 See, e.g., Taylor and Brown (1988), Weinstein (1980), Alloy and Abrahamson (1979) and the many other references given in section 3. For recent overviews of the general phenomenon of self-deception, see Gilbert and Cooper (1985) and especially Baumeister (1998) on the psychological evidence, Elster (1999) and Mele (1999) for the philosophical debates and implications.

6 These last two terms are borrowed from Seligman (1975, 1990).

7 Note that this last observation cannot readily be accounted for by the 'signalling' theory of self-confidence either.

8 The simplest date-0 action is thus the choice of the amount of information that will be available at date 1 (e.g., soliciting feedback, taking a test, keeping or destroying records). Alternatively, this information may be derived from the outcome of some activity pursued for its own sake at date 0 (learning by doing, drinking a lot of wine).

9 See Ainslie (1992, 2001) for the evidence, and Strotz (1956), Phelps and Pollack (1968), Loewenstein and Prelec (1992), Laibson (1997) and O'Donoghue and Rabin (1999) for formal models and economic implications.

10 Note that the equality in (2) makes use of the identity $\mathsf{E}_0[\theta|\mathsf{E}_1[\theta] = \bar{\theta}_1] = \bar{\theta}_1$, which holds when, and only when, there is no information loss between dates 0 and 1.

11 Rewrite (6) as $\mathcal{I}_F = \int_0^1 1_{\{\theta \geq c/\beta\delta v\}}(\delta\theta V - c)dF(\theta)$, where $1_{\{\cdot\}}$ denotes the indicator function, and note that the integrand is increasing in θ.

12 See, e.g., Berglas and Jones (1978), Arkin and Baumgardner (1985), Fingarette (1985) or Gilovich (199l).

13 See, e.g., Berglas and Baumeister (1993). Of course, self-handicapping involves both intrapersonal (self-esteem maintenance) and intrapersonal (self-presentation) motives; our model captures only the former. As Baumeister (1998) notes, 'by self-handicapping, one can forestall the drawing of unflattering attributions about oneself. Self-handicapping makes failure meaningless, and so if people think you are intelligent the upcoming test cannot change this impression.' In particular, people apparently self-handicap more in public situations (Kolditz and Arkin, 1982). They then reap a double dividend, as this provides an excuse for poor performance both to themselves and to others.

14 A third consideration is whether the expected cost of self-handicapping rises or falls with initial self-confidence: $h_0(F) \gtrless h_0(G)$. In Bénabou and Tirole (2000) we provide examples of tasks that correspond to each case.

15 Alternatively, it could arise from an affective or signalling value of self-esteem.

16 By definition, 'forgetting' means that an actual *loss of information* (a coarsening of the informational partition) occurs. Thus, if at date 1 the individual does not remember a date-zero signal σ, he also does not recall any other piece of information that is perfectly correlated with σ, such as the costs $M(\lambda)$ incurred in the process of forgetting. This complete forgetting is only a simplified representation of a richer attribution (signal extraction) problem, however. Let d measure

the level of an action that affects recall but can also be undertaken for its own sake: amount of wine consumed, time spent with friendly rather than critical people, attention paid to the details of competing informations, effort in making, safekeeping or disposing of physical records, spatial or mental detours around certain potential cues, etc. Choosing d following an event σ leads to a recall probability $\lambda = \Lambda(d)$ and has a direct utility $u_0(d, \varepsilon)$, where ε is a random taste shock. Later on, the agent may recall the action $d^*(\sigma, \varepsilon), \in \arg\max_d \{u_0(d, \varepsilon) + \beta\delta \mathsf{E}_0[u_1 + \delta u_2 \mid d, \sigma]\}$ that he took, and possibly the associated consequences $u_0(d^*(\sigma, \varepsilon), \varepsilon)$, but not the particular realization of ε that occurred (e.g., how much did I really want to drink wine, or sit next to that person at dinner?). Recalling $d^*(\sigma, \varepsilon)$ is thus generally insufficient to fully reconstruct σ, or to separate out within realized utility the cost $M(\Lambda(d^*(\sigma, \varepsilon), \varepsilon)) \equiv \max_d u_0(d, \varepsilon) - u_0(d^*(\sigma, \varepsilon), \varepsilon)$ that was incurred purely for memory manipulation. This problem remains when the functions u_0, Λ or M, or the distribution of ε, depend on σ.

17 For evidence and discussions see, e.g., Schacter (1996) and Fazio and Zanna (1981). Mullainathan (1999) and Laibson (2001) provide models of cue-dependent consumption.

18 Thus 'we can expect [an author in a meeting] to spend more time considering the comments of the lone dissenter who praised the project (and confirmed his self-conception) than of the colleagues who disliked it, thus mercifully softening the cavalcade of criticisms' (Gilbert and Cooper, 1985). For a discussion of self-presentation strategies and their link with self-enhancement, see Rhodewalt (1986).

19 One could even adhere to a minimalist version of the model where the individual can only improve his rate of recall (through rehearsal, record-making, etc.), but never lower it ($M(\lambda) = \infty$ for all $\lambda < \lambda_N$). All that matters is the potential for a *differential rate of recall* or awareness in response to desirable or undesirable informations.

20 As Gilbert and Cooper (1995) note, 'we are all insightful naive psychologists, well aware of human tendencies to be self-serving.'

21 In the first section we took the distribution of θ to be continuous, and c was fixed. In this section c has a continuous distribution and θ can take only two values. The two formulations are actually isomorphic (even if the latter happens to be more convenient here): all that really matters is the distribution of $\delta\theta V/c$.

22 While we assume here that λ can be freely varied between 0 and 1, the results would be identical if it were constrained to lie in some interval $[\underline{\lambda}, \bar{\lambda}]$. With $\underline{\lambda} > 0$ one can never forget (or avoid undesired cues) for sure.

23 It is also interesting to note that the specification with uniformly distributed costs is formally equivalent to one where c is fixed (say, $c \equiv 1$) but effort is a continuous decision, with net discounted payoff $\beta\delta\theta e - e^2/2$ for Self 1 and $\beta\delta(\delta\theta e - e^2/2)$ for Self 0. Thus, in our model, discontinuities in behaviour are not predicated on an indivisibility.

24 The terminology of 'beliefs and make-beliefs' is borrowed from Ainslie (2001).

25 Without denying the validity of this kind of evidence, we would like to emphasize that it should be interpreted with caution. Answers to surveys or experimental questionnaires may reflect self-presentation motives (for the benefits of the interviewer), or selective memory rehearsal strategies (for the individual's own benefit, as predicted by our model). Second, for every person who is 'overconfident' about how great they are (professionally, intellectually, socially, maritally), another one may be found who is underconfident, depressed, paralyzed by guilt and self-doubt,

but unlikely to acknowledge this to anyone except his closest confidant, counsellor or therapist. These could even be the same people at different points in time.

26 The interesting distinction is between self-perceived ability $E[\theta|\hat{\sigma}]$ and objectively assessed ability $E[\theta|\sigma]$, rather than between $E[\theta/\sigma]$ and the individual's true θ which it may measure only imperfectly. To simplify the exposition we shall therefore assume in this section that σ is perfectly informative about θ, i.e., $\theta = E[\theta|\sigma] \in \{\theta_L, \theta_H\}$. Alternatively, one could just read 'objectively assessed ability' wherever 'ability' appears.

27 Either this is the unique equilibrium, as in Figure 2.3, or else we focus on a symmetric situation for simplicity.

28 Note that these statements (like most experimental data) are about the agent's perception of his rank in the distribution of true abilities – not in the distribution of self-assessments which, as a Bayesian, he realizes are generally overoptimistic.

29 This was first pointed out by Carrillo and Mariotti (2000) for strategic ignorance, and is a feature that our model also shares with those of Brocas and Carrillo (1999) and Köszegi (1999).

30 There is of course a huge industry based on that premise, with countless web sites devoted to 'self-esteem', and hundreds of books with titles such as: '*How to Raise Your Self-Esteem', '31 Days to High Self-Esteem: How to Change Your Life So You Have Joy, Bliss & Abundance', '365 Ways to Build Your Child's Self-Esteem', '501 Ways to Boost Your Child's Self-Esteem', '611 Ways to Boost Your Self Esteem*: *Accept Your Love Handles and Everything About Yourself', 'ABC I Like Me'*, etc.

31 The intuition is relatively simple, however. The net loss across states from a 'hear no evil see no evil' strategy $\lambda^* = 0$, namely $-\Delta W(q, 0)$, is simply the *ex ante* value of information (always recalling the true σ, rather than having only the uninformed prior $\theta(0) = q$). Only when time inconsistency is strong enough can this value be negative.

32 This case is also interesting because it involves *two degrees of lack of commitment*: it is because the agent cannot commit to working at date 1 that his inability to commit not to tamper with memory at date 0 becomes an issue, which may end up hurting him more than if he had simply resigned himself to the original time-consistency problem.

33 For instance, an individual with $\beta < \bar{\beta}$ in Proposition 2 may be worse off when memory management is free ($M = 0$) than when it is prohibitively costly ($M(\lambda) = +\infty$ for all $\lambda \neq 1$). For instance, specification (a) above shows that such is always the case when the cost distribution $\varphi(c)$ is uniform.

34 Term $M(\mathrm{l})$ in (20)–(2l) is simply replaced by $\beta^{-1}\delta^{-\tau}\bar{M}/(1-q) + m(1)$, where $\bar{M}$ is the up-front cost of the commitment mechanism, $-\tau < 0$ is the period when the commitment was made, and $m(1) \geq 0$ is the cost of perfect recall faced at $t = 0$ as a result of this decision (whereas $m(\lambda) = +\infty$ for all $\lambda < 1$).

35 This is the kind of example to which Carrillo and Mariotti (2000) apply their model of strategic ignorance, pointing to studies that suggest that most people actually overestimate the health risks from smoking. More generally, the role of the timing of costs and benefits is emphasized in Brocas and Carrillo (2000).

36 A more general formulation, encompassing both affective and instrumental concerns, would be $\mathsf{E}_0[\max\{\mathsf{E}_1[\theta V - c], 0\} + J(F_1)]$, where $F_1(\theta)$ denotes the agent's date-0 and date-1 subjective probability distribution over his true ability. The functional $J(\cdot)$ represents either an exogenous hedonic utility, or an endogenous value function capturing the instrumental value of beliefs for self-motivation or self-presentation (signalling) purposes. In our model $J(F_1)$ is easily computed, and related to β.

37 In a different context, Rabin (1995) makes beliefs about the negative externalities of one's actions (on other people, animals, or the environment) an argument of the utility function, and assumes concavity. This provides an explanation for why people may prefer not to know of the potential harm caused by their consumption choices. Caplin and Leahy (2001) study a general class of preferences where initial perceptions of future lotteries enter into the intertemporal utility function. Depending on whether the dependence is concave or convex, a person will choose to avoid information that would make the future lottery more risky or, on the contrary, seek out information and situations that increase the stakes.

References

Ainslie, G. (1992) *Picoeconomics: The Strategic Interaction of Successive Motivational States Within the Person*. Studies in Rationality and Social Change (Cambridge, England: Cambridge University Press).

Ainslie, G. (2001) *Breakdown of Will* (Cambridge: Cambridge University Press).

Akerlof, G. and W. Dickens (1982) 'The Economic Consequences of Cognitive Dissonance', *American Economic Review*, vol. 72, pp. 307–19.

Alloy, L.T. and L. Abrahamson (1979) 'Judgement of Contingency in Depressed and Nondepressed Students: Sadder but Wiser?' *Journal of Experimental Psychology: General*, vol. 108, pp. 441–85.

Arkin, R.M. and A.H. Baumgardner (1985) 'Self-Handicapping', in J. Harvey and G. Weary (eds), *Attribution: Basic Issues and Applications* (New York: Academic Press), pp. 169–202.

Bandura, A. (1977) *Self Efficacy: The Exercise of Control* (New York, NY: W.H. Freeman).

Baumeister, R. (1998) 'The Self', in D. Gilbert, S. Fiske and G. Lindzey (eds), *The Handbook of Social Psychology* (Boston: McGraw-Hill), pp. 680–740.

Bénabou, R. and J. Tirole (1999) 'Self-Confidence: Intrapersonal Strategies', IDEI-Université de Toulouse, mimeo, June.

Bénabou, R. and J. Tirole (2003) 'Intrinsic and Extrinsic Motivation', *Review of Economic Studies*, vol. 70, pp. 489–520.

Bénabou, R. and J. Tirole (2004) 'Willpower and Personal Rules', *Journal of Political Economy*, vol. 112, pp. 848–87.

Berglas, S. and E. Jones (1978) 'Drug Choice as a Self-Handicapping Strategy in Response to Non-Contingent Success', *Journal of Personality and Social Psychology*, vol. 36, pp. 405–17.

Brocas, I. and J. Carrillo (2004) 'Entrepreneurial Boldness and Excessive Investment', *Journal of Economics and Management Strategy*, vol. 13, pp. 321–50.

Brocas, I. and J. Carrillo (2000) 'The Value of Information when Preferences are Dynamically Inconsistent', *European Economic Review*, vol. 44, pp. 1104-15.

Caplin, A. and J. Leahy (2001) 'Psychological Expected Utility Theory and Anticipatory Feelings', *Quarterly Journal of Economics*, vol. 116, pp. 55–79.

Carrillo, J. and T. Mariotti (2000) 'Strategic Ignorance as a Self-Disciplining Device', *Review of Economic Studies*, vol. 66, pp. 529–44.

Crary, W.G. (1966) 'Reactions to Incongruent Self-Experiments', *Journal of Consulting Psychology*, vol. 30, pp. 246–52.

Darley, J. and G. Goethals (1980) 'People's Analyzes of the Causes of Ability-Linked Performances', in L. Berkowicz (ed.), *Advances in Experimental Social Psychology*, vol. 13 (New York: Academic Press), pp. 1–37.

Deci, E. (1975) *Intrinsic Motivation* (New York: Plenum Press).

Elster, J. (1999) 'Motivated Belief Formation', Columbia University mimeo, June.
Fazio, R., and M. Zanna (1981) 'Direct Evidence and Attitude-Behavior Consistency', in L. Berkowitz (ed.) *Advances in Experimental Social Psychology*, vol. 14 (New York: Academic Press), pp. 162–98.
Festinger, L. (1954) 'A Theory of Social Comparison Processes', *Human Relations*, vol. 7, pp. 117–40.
Fingarette, H. (1985) 'Alcoholism and Self-Deception', in M. Martin (ed.), *Self-Deception and Self-Understanding* (Lawrence, KS: University Press of Kansas), pp. 52–67.
Freud, S. (1938) *A General Introduction to Psychoanalysis* (New York: Garden City Publishing Co.).
Frey, D. (1981) 'The Effect of Negative Feedback about Oneself and Cost of Information on Preference for Information about the Source of this Feedback', *Journal of Experimental Social Psychology*, vol. 17, pp. 42–50.
Gilbert, D. and J. Cooper (1985) 'Social Psychological Strategies of Self- Deception', in M. Martin (ed.), *Self-Deception and Self-Understanding* (Lawrence, KS: University Press of Kansas), pp. 75–94.
Gilovich, T. (1991) *How We Know What Isn't So* (New York: Free Press).
Greenier, K., M. Kernis, and S. Wasschul (1995) 'Not All High (or Low) Self-Esteem People Are the Same: Theory and Research on the Stability of Self-Esteem', in M. Kernis (ed.), *Efficacy, Agency and Self-Esteem* (New York: Plenum Press), pp. 51–72.
Greenwald, A. (1980) 'The Totalitarian Ego: Fabrication and Revision of Personal History', *American Psychology*, vol. 35, pp. 603–13.
Gur, R. and H. Sackeim (1979) 'Self-Deception: A Concept in Search of a Phenomenon', *Journal of Personality and Social Psychology*, vol. 37, pp. 147–69.
Heider, F. (1958) *The Psychology of Interpersonal Relations* (New York: Wiley).
James, W. (1890) *The Principles of Psychology* (Cleveland, OH: World Publishing).
Jones, E., F. Rhodewalt, S. Berglas, and J. Skelton (1981) 'Effects of Strategic Self-Presentation on Subsequent Self-Esteem', *Journal of Personality and Social Psychology*, vol. 40, pp. 407–21.
Kolditz, T. and R. Arkin (1982) 'An Impression-Management Interpretation of the Self-Handicapping Strategy', *Journal of Personality and Social Psychology*, vol. 43, pp. 492–502.
Korner, I. (1950) *Experimental Investigation of Some Aspects of the Problem of Repression: Repressive Forgetting*. Contributions to Education, no. 970 (New York, NY: Bureau of Publications, Teachers' College, Columbia University).
Köszegi, B. (1999) 'Self-Image and Economic Behavior', MIT mimeo, October.
Kunda, Z. and R. Sanitioso (1989) 'Motivated Changes in the Self-Concept', *Journal of Personality and Social Psychology*, vol. 41, pp. 884–97.
Laibson, D. (1997) 'Golden Eggs and Hyperbolic Discounting', *Quarterly Journal of Economics*, vol. 112, pp. 443–78.
Laibson, D. (2001) 'A Cue-Theory of Consumption', *Quarterly Journal of Economics*, vol. 116, pp. 81–119.
Laughlin, H.P. (1979) *The Ego and Its Defenses*. The National Psychiatric Endowment Fund eds., second edition (New York, NY: Jason Aaronson, Inc.).
Leary, M. and D. Down (1995) 'Interpersonal Functions of the Self-Esteem Motive: The Self-Esteem System as Sociometer', in M. Kernis (ed.), *Efficacy, Agency and Self-Esteem* (New York: Plenum Press), pp. 123–44.
Loewenstein, G. and D. Prelec (1992) 'Anomalies in Intertemporal Choice: Evidence and Interpretation', *Quarterly Journal of Economics*, vol. 107, pp. 573–97.

Mele, A. (1999) 'Real Self-Deception', *Behavioral and Brain Sciences*, vol. 20, pp. 91–136.
Mischel, W., E.B. Ebbesen, and A.R. Zeiss (1976) 'Determinants of Selective Memory about the Self', *Journal of Consulting and Clinical Psychology*, vol. 44, pp. 92–103.
Mullainathan, S. (2002) 'A Memory Based Model of Bounded Rationality', *Quarterly Journal of Economics*, vol. 117, pp. 735–74.
Murray, S.L. and J.G. Holmes (1994) 'Seeing Virtues in Faults: Negativity and the Transformation of Interpersonal Narratives in Close Relationships', *Journal of Personality and Social Psychology*, vol. 20, pp. 650–63.
Nisbett, R. and T. Wilson (1977) 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review*, vol. 84, pp. 231–59.
O'Donoghue. T. and M. Rabin (1999) 'Doing it Now or Later', *American Economic Review*, vol. 89, pp. 103–24.
Phelps, E. and R. Pollack (1968) 'On Second-Best National Savings and Game-Equilibrium Growth', *Review of Economic Studies*, vol. 35, pp. 185–99.
Rabin, M. (1995) 'Moral Preferences, Moral Rules, and Belief Manipulation', University of California mimeo, April.
Rhodewalt, F.T. (1986) 'Self-Presentation and the Phenomenal Self: On the Stability and Malleability of Self-Conceptions', in R. Baumeister (ed.), *Public Self and Private Self* (New York: Springer Verlag), pp. 117–42.
Salancik, G. (1977) 'Commitment and the Control of Organizational Behavior and Belief', in B. Staw and G. Salancik (eds.), *New Directions in Organizational Behavior* (Chicago: St. Clair Press), pp. 1–54.
Sartre, J.P. (1953) *The Existential Psychoanalysis* (H.E. Barnes, trans.) (New York: Philosophical Library).
Schacter, D. (1996) *Searching for Memory* (New York: Basic Books).
Seligman, E. (1990) *Learned Optimism: How to Change Your Mind and Your Life* (New York: Simon and Schuster).
Snyder, C. (1985) 'Collaborative Companions: The Relationship of Self-Deception and Excuse Making', in M. Martin (ed.), *Self-Deception and Self-Understanding* (Lawrence, KS: University Press of Kansas), pp. 35–51.
Strotz, R. (1956) 'Myopia and Inconsistency in Dynamic Utility Maximization', *Review of Economic Studies*, vol. 22, pp. 165–80.
Swann, W.B. Jr. (1996) *Self Traps: the Elusive Quest for Higher Self-Esteem* (New York: W.H. Freeman and Company).
Taylor, S.E. and J.D. Brown (1988) 'Illusion and Well-Being: A Social Psychological Perspective on Mental Health', *Psychological Bulletin*, vol. 103, pp. 193–210.
Weinberg, B. (1999) 'A Model of Overconfidence', Ohio State University mimeo, August.
Weinstein, N. (1980) 'Unrealistic Optimism About Future Life Events', *Journal of Personality and Psychology*, vol. 39, pp. 806–20.
Zuckerman, M. (1979) 'Attribution of Success and Failure Revisited, or the Motivational Bias is Alive and Well in Attribution Theory', *Journal of Personality*, vol. 47, pp. 245–87.

3

Rationality, Learning and Complexity*

Alessandro Vercelli
University of Siena, Italy

1 Introduction

Although standard economic theory is based on methodological individualism, this does not imply that individuals play a crucial role in economic models. On the contrary, in such a theory individuals are deprived of authentic subjective features and play no significant role as genuine subjects. The so-called *homo economicus* is characterized by given preferences that are conceived as exogenous and invariant over time. Therefore, the genuine psychological features of an economic agent do not matter.

This chapter focuses on the impact of cognitive psychology on economic behaviour. Casual observation and experimental research suggest that cognitive psychology significantly affects expectations and learning, which in turn play a crucial role in economic decisions. In standard economics, however, expectations and learning are conceived in such a way that cognitive psychology becomes irrelevant. We intend to clarify the reasons for this neglect and to specify the conditions under which the chasm between economics and cognitive psychology may be reduced. The crucial obstacle to closing this gap is the very narrow notion of rationality entertained by standard economics – generally called 'substantive rationality' (Simon, 1976) – which

* A preliminary version of the ideas discussed in this chapter was presented at the conference 'Keynes, Knowledge and Uncertainty' (University of Leeds, March 1996) and published in its proceedings (Vercelli, 2002). Successive versions have been presented on many occasions, including the annual conference of Anpac (San Salvador de Bahia, December 2001), the annual conference of EAEPE (Siena, November 2001), the first meeting of the Italian Society of Cognitive Sciences (Rovereto, September 2002), the World Congress of IEA (Lisbon, September 2002), and the conference on 'Complex Behaviour in Economics' (Aix-en-Provence, May 2003). I thank the audience at each of my presentations for their constructive comments. In addition, I wish to thank Bina Agarwal, Samuel Bowles and Stefano Fiori for their helpful suggestions.

implies restrictive notions of expectation-formation and learning that deny a role to cognitive psychology.[1]

Substantive rationality is regarded as a suffocating straitjacket by an increasing number of economists, many of whom are actively exploring new research avenues. There are basically two strategies in such an exploration: the progressive relaxation of some of the most restrictive assumptions of the standard approach to make it more generally applicable, and the search for a more satisfactory alternative approach.[2] These endeavours have led to many important insights in various specialized subfields of economic theory that go beyond the narrow limits of substantive rationality.[3] However, these new insights still need to be integrated into a fully-fledged economic paradigm that can replace the traditional, substantive-rationality paradigm. With this goal in mind, this chapter aims to individuate the main assumptions underlying the traditional paradigm and to classify the deviations from it in coherent alternative paradigms. The conceptual backbone of our argument is thus based on the taxonomy of some basic concepts in economics – rationality, learning, expectations, uncertainty – and a few ontological assumptions about the nature of time and of the 'World', that is, the empirical evidence under investigation. The chapter shows that between these concepts there are relationships of semantic congruence or incongruence, with interpretive and normative implications. In particular, we intend to argue that it is impossible to rely on the standard paradigm to take into account serious complexity of economic behaviour. Alternative paradigms are required, the characteristics of which we aim to clarify in the following sections.

The definitions of complexity vary according to the epistemological and ontological context of the analysis. Because the scope of this chapter is rather broad, our argument is not based on a unifying definition of complexity. However, there is a clear relationship between the concept of complexity underlying the taxonomy mentioned above, and the standard measures of complexity. The latter are syntactic or computational orderings concerning the minimal number of logical steps or amount of information necessary to describe a certain system or to determine a particular class of transformations. The most popular measures of complexity refer to properties of linguistic systems, dynamic models and computing machines. (For a critical survey with particular reference to economics, see Foley, 2001, and Albin, 2001.) In each of these cases, four basic levels of complexity are distinguished; the levels in one ordering are strictly correlated with the corresponding levels in the other orderings. From a dynamic perspective, Level 1 (minimal complexity) corresponds to stable linear systems with a unique equilibrium; Level 2 corresponds to stable systems that have a regular periodic attractor, such as a limit cycle; Level 3 corresponds to non-linear systems characterized by chaotic behaviour and monotonic propagation of disturbances; and Level 4 corresponds to chaotic behaviour and irregular propagation of disturbances (*ibid.*). In this chapter, however, complexity

is mainly discussed from a semantic point of view; that is, with respect to the congruence between the degree of complexity of the World and that of its subjective representations. This semantic approach makes it clear that standard economics is applicable, under further limiting conditions, only if the degree of complexity of the empirical evidence under examination does not exceed Level 2. To take account of higher levels of complexity, alternative approaches based on more sophisticated concepts of rationality and learning are required.

The structure of the chapter is as follows. Section 2 introduces the basic framework of the analysis, centred on the interaction between subject and object in economics. Such interaction may be characterized by different concepts of rationality, learning and expectations formation. In this section, the usual specification of standard economics – based on substantive rationality – is clarified in terms of its inner logic. In section 3, more general concepts of rationality are introduced (procedural and behavioural rationality), to overcome the strictures of substantive rationality. In section 4, the relationship between expectations, learning and rationality is set out in some detail. In section 5, more general concepts of learning are introduced to accommodate the more general concepts of rationality and the economic value of those concepts of learning is assessed and compared. In section 6, the implications of various ontological assumptions concerning the nature of uncertainty, the irreversibility of time, and the nature of the World are clarified. In section 7, the correspondence between the different versions of the key concepts examined in the preceding sections are discussed and their normative implications are made explicit. The chapter closes with some brief remarks.

2 The interaction between subject and object and the strictures of substantive rationality

A subject may be defined only in relation to an object. Therefore, an analysis of the subjective features of economic behaviour must start from a representation of the interaction between subject and object in economics; that is, the interaction between economic agent and economic system (which, for the sake of brevity, is referred to as 'the Interaction' in this chapter). According to the usual conceptualization of the Interaction, an economic agent is characterized by (1) a set of preferences, (2) an initial endowment of resources, and (3) a set of options, each of which leads to well-specified outcomes for each possible state of the World. Whenever economic agents face an intertemporal decision problem, they choose the best possible option according to their preferences and expectations of how the key variables will affect the value of the objective function in the future. These expectations are based on a 'model'; that is, on as accurate a representation as possible of the part of the World that is relevant for the decision problem to be faced.

In this chapter, the epistemic and ontological properties of the Interaction are sharply distinguished to study their mutual influence. The ontological properties of the World are defined as properties of the economic system that may be considered 'true', regardless of their epistemic representation; the epistemic properties of the model are provisionally assumed by an agent as a representation of the corresponding ontological properties, but are likely to be revised and corrected to represent them better. The divergence between an ontological property and its subjective representation is defined here as 'stochastic error' when it depends exclusively on exogenous shocks, and 'systematic error' when it depends on some intrinsic bias of the representation; for example, a different value attributed to one parameter of the probability distribution of a key variable. Within this conceptual framework, learning is defined as the epistemic process aimed at reducing, and possibly eliminating, the systematic errors, while rationality may be defined essentially as the ability to learn. The feedback between rationality and learning is crucial in determining the economic behaviour of decision-makers, and is therefore central to the argument that follows. The Interaction, as represented here, is deceptively simple, since it involves a crucial self-referential loop between subject and object. The behaviour of economic agents is part of the World and is affected by its subjective representations and expectations (Figure 3.1).[4] We wish to emphasize that the way

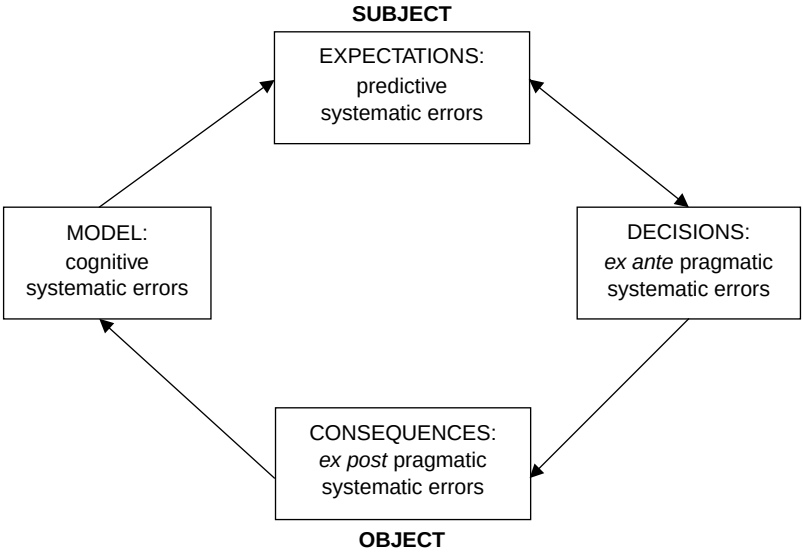


Figure 3.1 The interaction between subject and object

in which the Interaction is conceived and modelled is crucial in classifying economic theories and assessing their capability to represent, predict and control complex economic behaviour.

The dynamic decision problem, as defined above, may give determinate results only by resorting to simplifying assumptions. In standard economics this role is played by the assumption that the agent is rational from both the epistemic (cognitive) and pragmatic (practical) points of view, in a specific sense that must be accurately defined and modelled. According to the standard view of substantive rationality, epistemic rationality is implicitly specified as the ability to avoid systematic errors. In a deterministic decision problem, this implies that an agent has perfect foresight. In a stochastic decision problem, it implies that the agent has rational expectations that, by definition, avoid any sort of systematic error.[5] In a deterministic model, the expected value is the correct one. In a stochastic model, the subjective probability distribution of the prediction is assumed to coincide with the 'objective' or 'true' probability distribution (see, for example, Minford-Peel, 1983). As for pragmatic rationality, it is assumed that no systematic errors are made in the execution of the decision, so that the maximization *ex ante* of the objective function translates to its effective maximization *ex post*.

In this view, the subjective characteristics of an economic agent are irrelevant, since the model – and therefore the expectations based on it – is by definition 'true' in the sense that the subjective representation of the world 'coincides' with its objective features. The preferences of a decision-maker play a crucial role in the choice of the option that maximizes the objective function, but this does not involve any role for the psychological characteristics of the agent. Such preferences should express, by definition, the subjective features of an individual – that is, his or her tastes – but in the standard approach they are exogenous and given. Moreover, self-interest is the only motivation taken into account, completely neglecting social motivations such as altruism, equity and solidarity, whose role in the real world is confirmed by massive empirical evidence and extensive experimental work (for example Kahneman, 2003).

Cognitive psychology has no role to play within the assumptions of substantive rationality, since economic agents are assumed to base their choices on the 'true' model of the real world. Learning does not completely disappear, but in this context it retains only a trivial role: the updating of the information-set as soon as new observed values of the relevant variables occur. However, this role does not involve the influence of the genuine subjective cognitive features of the agent, because the same role could be played by a computer that collects all the relevant news in real time to update the information-set that is relevant for the decision problem, exactly as a seismograph records data on earthquakes. We may conclude that, within the framework of standard economics, the psychological and cognitive features of agents are neglected altogether.

Under the assumptions of substantive rationality, the Interaction is thus described by a dynamic process that is assumed to be in equilibrium. In the absence of systematic errors *ex post*, there is no reason to revise the model of the World and the expectations based on it. Since the axiom of intertemporal coherence (that is crucial in the definition of substantive rationality) ensures that the preferences of an agent do not change over time, the agent makes the same choices under the same circumstances, and the Interaction settles into a stationary equilibrium.

3 More general concepts of rationality

The advantage of substantive rationality, on the basis of its founding axioms, is that equilibrium is taken for granted in the Interaction, and therefore in the economic system. This greatly simplifies the analysis of economic behaviour, permitting the systematic application of sophisticated formalization, leading to determinate results. The trouble is that the same axioms restrict the scope of legitimate application to a closed, simple and 'familiar' World; that is, to problems that are well-known and characterized by stationarity, trivial uncertainty and the irrelevance of the psychological and cognitive features of the agent. This was recognized long ago by economists who were dissatisfied with the standard approach. Schumpeter (1934) explained that the standard approach may be legitimately applied only to the 'circular flow' characterized by routine economic behaviour, but cannot be applied to the process of development, since such a process is characterized by non-stationarity, innovation and radical uncertainty. Keynes (1936) contended that the axioms of the standard approach are like those of Euclidean geometry that cannot be legitimately applied to a non-Euclidean world; that is, whenever hard uncertainty and structural instability play a crucial role, as is typical in a monetary economy. Simon (1976), as we have seen, described the crucial role played by the assumption of 'substantive rationality' in narrowing the empirical scope of standard economics, and suggested a more general concept – 'procedural rationality' – borrowed from cognitive psychology.

The growing awareness of the strictures of the traditional paradigm of substantive rationality has led many economists to conceive of more general concepts of rationality.[6] They share a sharp rebuttal of the unlimited power of substantive rationality. In particular they question the alleged ability of economic agents always to entertain the 'true' representation of the world, to formulate correct forecasts of the relevant variables (apart from intrinsically unpredictable exogenous shocks), to choose the best available option, and to implement it in the most efficient way. In this view, rationality is characterized by unavoidable bounds of cognitive, computational, psychological and sociological nature. The contributions of these economists, thus, are often referred to under the heading of 'bounded rationality'. This broad category, however, is too generic and blurs the deep differences between

the various concepts of rationality that are critical of substantive rationality. In particular, in different approaches, different bounds are emphasized and conceived as more or less limiting. In addition, the relationship between the positive and normative sides of rationality is conceived differently in alternative versions of the bounded rationality approach. Here, it is sufficient to distinguish between two broad varieties of bounded rationality: 'procedural rationality', as defined by Simon (1976), who was the first to speak of the bounded nature of rationality; and what we suggest be referred to as 'behavioural rationality', a category that groups more general concepts of bounded rationality that further relax its axioms. In particular, 'behavioural rationality' includes the contributions of March (1988), Kahneman (2003) and other followers of Simon, who generalized his insights on the descriptive features of the effective rationality of economic agents, under the stimulus of systematic observation and laboratory experiments.

The concept of procedural rationality permits the analysis of the Interaction from the point of view of disequilibrium. Whenever a systematic *ex post* error is detected by an agent, the model of the World is revised to correct the expectations and to eliminate the error, or at least to reduce its size. From this more general viewpoint, the equilibrium of the process does not need to imply a complete absence of systematic errors. According to the criterion suggested by Simon, the dynamic process of learning may stop as soon as the size of the errors is small enough to be 'satisficing' relative to the aspiration level of an agent (Simon, 1982: 415). A further attempt to eliminate the residual errors, in a situation characterized by limited information and bounded rationality, could imply greater economic and psychological costs than benefits. When conceived in this way, the dynamic process of learning does not need to be linear, and may be complex. In particular, the equilibrium may depend on the initial conditions and the path followed. In this view, substantive rationality may be seen as a special case of procedural rationality, because it focuses exclusively on a particular equilibrium characterized by the absence of systematic errors to which the process of learning could converge under a series of restrictive assumptions. The principal assumptions are: linearity of the process of learning; no cost of transactions and of learning; weak or 'soft' uncertainty (section 6), and unlimited cognitive and computational capabilities.

Simon's concept of procedural rationality was heavily influenced by nineteenth-century logicism (Frege, Russell and Whitehead) and by the emerging Artificial Intelligence paradigm that led Simon to conceive of rationality in terms of explicit and logically consistent algorithmic procedures. Intuition was conceived as mere pattern recognition, drawn from long-term memory, and creativity as a mere recombination of preexisting elements. What we designated 'behavioural rationality' relaxes all of these limitations; it calls attention beyond 'algorithmic' cognitivism to other subjective aspects of psychology (intuition, framing and emotions), and includes unconscious and inconsistent drives. This approach extends the

empirical scope of economic theory, but risks a weakening of the normative implications of rationality.

The positive and normative sides of rationality are strictly related in the traditional view of substantive rationality. In Simon's approach of procedural rationality, the link between these sides is still quite strong and is based on the concept of 'satisficing', which is related to the level of aspiration of economic agents. In behavioural rationality, the normative implications are not always clear, but in principle may overcome the limitations of alternative concepts of rationality. In fact, from the perspective of both substantive rationality and procedural rationality, the normative side of rationality refers to the process by which an agent adapts to a given environment. The fact that procedural rationality applies also to the disequilibrium dynamics of the Interaction allows a proper understanding of the process of adaptation to an environment that is conceived as given, although possibly open. In this view, agents are option-takers in the sense that they cannot alter the set of options with which they are initially endowed. Evolution theory, however, tells us that *homo sapiens* adapted to the environment by modifying the environment to their needs by creating new options. This was progressively developed over the course of human evolution to degrees unimaginable for other living beings, including the other primates. While animal rationality was purely adaptive, *homo sapiens* developed a particular kind of rationality that became progressively more proactive. The reciprocal adaptation of humans to their environment was sought on the basis of a project or design that was updated or revised according to its success.[7] In behavioural rationality, agents are also creators of options since they are able to introduce new options through advances in scientific research and technology. Therefore, human rationality reacquires the specific features that distinguish our species from others, features that are completely neglected by substantive rationality and, to some extent, by standard procedural rationality.[8] The latter may be considered a limiting case of behavioural rationality when the set of options is given and invariant.

4 Expectations, learning and rationality

Returning to the Interaction, after implementing the chosen option, economic agents observe its consequences. Whenever they ascertain the absence of systematic errors *ex post* – as in the case of substantive rationality by definition – there is no reason to revise the model and the expectations based on it, apart from its necessary updating. In this case the Interaction remains in a state of equilibrium that may be shifted by exogenous factors as registered by the process of updating. In other words, substantive rationality focuses exclusively on the equilibrium path of the Interaction. This approach is reasonable and productive to the extent that we are able to provide dynamic foundations for it by showing that this equilibrium path is stable and the convergence to it is sufficiently rapid; otherwise, the

scope of the theory and of its applications would be very limited. For this reason, an extensive literature, stretching back to at least the study of the *tâtonnement* by Walras (1874), has investigated the stability of equilibrium. On the whole, the results have been quite discomforting: the stability of a Walrasian general equilibrium model requires quite demanding conditions;[9] crucially, the convergence towards the substantive-rationality equilibrium depends on the process of learning.

To clarify the relationship between rationality, learning and expectations, let us express their analytical relationship. We start from the deterministic case that was dominant in economic theory until the 1960s. In this case, the prevailing paradigm may be expressed through the following difference equation:

$$x_t - x_{t-1} = a[T(x_{t-1}) - x_{t-1}] + k_t \tag{1}$$

where x_t represents the expectations of a certain variable that describes the behaviour of the World, $T(x_t)$ designates the effective behaviour of such a variable that is affected by its expectations, and k_t is an exogenous variable that affects the dynamics of the endogenous variable and its expectations.[10] The dynamics of expectations, thus, depends on two terms. The first, $a[T(x_{t-1}) - x_{t-1}]$, describes the endogenous dynamics of the interaction between the subject who formulates the expectations and the economic system. The endogenous dynamics depends on what we call the systematic error observed *ex post*: $T(x_{t-1}) - x_{t-1}$. Whenever the endogenous dynamics is stable – that is, the systematic error tends to zero – we have genuine learning and convergence towards the equilibrium path of the Interaction. The exogenous dynamics determined by the second term, k_t, specifies the nature of the equilibrium path that is stationary when the exogenous term is stationary, and is mobile whenever the exogenous term is a function of time.

When the term $[T(x_{t-1}) - x_{t-1}]$, describing the endogenous dynamics of the interaction, is systematically zero, we have the case of perfect foresight that, in a deterministic environment, is by definition the only case fully consistent with substantive rationality. However, since perfect foresight is a very demanding assumption for applied economics, in the 1960s an alternative hypothesis of expectations formation, called the adaptive expectations hypothesis (AEH), became dominant (Cagan, 1956). The reason for its success rested on its capacity to deal, to some extent, with the existence of systematic errors and genuine learning. Notwithstanding the mathematical sophistication of this approach, the idea was simple and was constrained in such a way as to avoid challenging the paradigm of substantive rationality while adding realism to it. It was assumed that rational agents are able to learn so that the parameter a is always in the range of values that ensure dynamic stability. The endogenous part, moreover, was assumed to be relevant for

the short period (in reference to the business cycle) while the exogenous part, fully consistent with the axioms of substantive rationality, was considered the centre of attraction of the system that always prevails in the long period (in reference to growth). In addition, the linear specification in equation (1), which excludes complex dynamics, is crucial because it ensures that the equilibrium path is not affected by the disequilibrium dynamics. The substantive-rationality paradigm was thus somehow saved as the gravitational centre of the economy, while the introduction of disequilibrium dynamics was intended only to confer on it more realism.

The paradigm of adaptive expectations broke down at the end of the 1960s, for a series of reasons connected with the history of both facts (stagflation) and ideas. Here, we mention only briefly a few crucial reasons related to the history of economic thought. The AEH paradigm was fully consistent with the conviction – prevailing until the late 1960s – that economic equilibrium needed sound dynamic foundations. This was recognized by Walras (1874), who tried to solve the problem by studying the process of *tâtonnement*. An important breakthrough came with Samuelson's (1947) doctoral thesis which proposed for the first time comprehensive dynamic foundations for the general equilibrium theory. This approach was influential among many economists during the two decades that followed and clearly underlay the success of the AEH. In consequence of the axiomatization proposed by Debreu (1959), a new approach emerged in economic theory according to which general equilibrium theory needed only axiomatic foundations (Ingrao-Israel, 1990). In this approach, axiomatic foundations guarantee the logical consistency of the theory, while the empirical validity of the theory is assessed by statistic and econometric tests. For this reason, dynamic foundations seem superfluous. From the viewpoint of the axiomatic foundations of general equilibrium theory, the AEH is clearly unacceptable because it implies the existence of systematic errors, which is logically inconsistent with the axioms of the theory, in particular, the axiom of intertemporal coherence. Logically, the alleged small size and vanishing nature of errors are hardly acceptable excuses.

The same circle of economists and mathematicians elaborated a fully fledged theory of general equilibrium under uncertainty (Arrow, 1953; Arrow and Hahn, 1971). This permitted a stochastic interpretation of equation (1) that seemed to provide a better solution. The equilibrium was then described in terms of probability distributions of the relevant variables, so that it could allow for the possibility of *ex post* errors, provided that they were not systematic. This added to the apparent realism of the substantive-rationality interpretation of the Interaction. In the same years, a new hypothesis of expectations formations – the REH, rational expectations hypothesis (Muth, 1961) – was suggested in the literature, and turned out to be fully consistent with this view. According to the REH, it is assumed that rational agents can always avoid systematic errors so that their subjective probability

distributions of the relevant variables are assumed to coincide with their objective distribution (see for example Begg, 1982). In this view, there is no room for genuine learning. The only kind of learning consistent with – indeed necessary to – this view is the real-time updating of the information-set that ensures the persistent and continuous absence of systematic errors.

The issue of equilibrium stability, however, did not disappear from the debate, because if the logical possibility of disequilibrium is not excluded, it is necessary to know whether there are sound reasons for focusing the analysis exclusively on the equilibrium. To clarify this issue, we may refer to the small but insightful literature on the stability of the rational-expectations equilibrium.[11] Two basic streams can be distinguished, based on the study of either expectational stability or learning rules.

The first stream originated through the seminal contributions of Lucas (1978), DeCanio (1979) and Evans (1983). According to Evans and Honkapohja (1994), the analysis of expectational stability may be ultimately reduced to the following difference equation:

$$x_t - x_{t-1} = a[T(x_{t-1}) - x_{t-1}] \tag{2}$$

This is the same as equation (1), but without the exogenous factor that by definition does not affect the stability of the system. This equation is intended to describe a stylized process of learning that occurs in time t, which determines a progressive reduction in the gap between the dynamics of expectations x_t, and the effective dynamics $T(x_t)$, which is a function of perceived dynamics. The dynamics of expectations is thus a function of the systematic *ex post* errors, defined by $T(x_t) - x_t$. Actual learning implies a process of convergence towards the rational-expectations equilibrium. As soon as this equilibrium is effectively reached, the systematic errors vanish and the process of genuine learning stops.

The second stream focuses on the study of the learning rules in real time and is based on contributions by Bray (1982), Bray and Savin (1986), Fourgeaud, Gourieroux and Pradel (1986), Marcet and Sargent (1989a,b) and Woodford (1990). Learning rules are expressed in terms of approximation algorithms: in particular, recursive least squares and recursive ARMA estimations. In this case, the learning process may be expressed through a dynamic equation that is a function of the systematic *ex post* errors. For example, the seminal model by Bray (1982) may be expressed by the following stochastic approximation algorithm:

$$\beta_t = \beta_{t-1} + (1/t)[p_{t-1} - \beta_{t-1}] \tag{3}$$

where p_t is the effective price at time t, and β_t is the expectation of p_{t+1}, which is equal – by hypothesis – to the average of realized prices. Here,

the dynamic behaviour of expectations depends exclusively on the *ex post* systematic errors, which are expressed in equation (3) by the term inside the square brackets (that is, by the deviation of the latest observation from the average of the past values). The process of learning stops only if the *ex post* systematic errors are fully corrected; that is as soon as the rational expectations equilibrium is reached (Sargent, 1993: 88 note 2). This confirms that within the REH, economic agents are not allowed to make systematic errors *ex post*: therefore genuine learning – that is the correction of systematic errors – is inconceivable in any theory or model based on this hypothesis. On the contrary, convergence towards the rational-expectations equilibrium, which can be considered as a process of genuine learning, would imply that while agents are learning, they are not allowed to form expectations based on the REH. Therefore, irrespective of whether or not the learning process converges towards a rational expectations equilibrium, this literature proves the inconsistency between rational expectations and genuine learning. The only kind of learning that is truly consistent with the REH is the trivial real-time updating of the relevant stochastic variables, which by hypothesis does not affect the parameters of those processes.

The way out of this dilemma can only be found in alternative hypotheses of expectations formation that do not rule out the ubiquity of systematic errors and the crucial role of genuine learning in correcting them. To proceed in this direction, the tradition of the AEH may be rescued and consistently developed within the assumptions of broader concepts of rationality. This may be achieved by eliminating the constraints introduced to comply with the axioms of substantive rationality. In this different framework, the criticisms levelled against adaptive expectations from the viewpoint of substantive rationality lose their logical strength. In particular, the resting point of the dynamic system need not be unique, may be path-dependent, and could be reached in finite time, as soon as the perceived marginal costs of learning – plus transition costs – exceed its perceived marginal advantages. In this case, the possible persistence of systematic errors becomes fully consistent with economic logic. In addition, the process of revising expectations does not need to be backward-looking, as it does in the traditional specification of the AEH. The announcement of a change in the policy environment leads bounded-rational agents to redefine rather rapidly the resting points of the system that comply with the criterion of satisficing. This may cause a shift in behaviour, similar to that occurring in the REH, but is unlikely to eliminate the systematic errors or to stop the process of genuine learning. Finally, from the perspective of behavioural rationality, expectations may be proactive in the sense that their influence on future events may be taken into account to realize a 'creative' adaptation to the environment, modified according to a rational design.

5 Concepts of learning and their economic value

In the light of the preceding analysis, we may investigate the economic value of the different concepts of learning within the different concepts of rationality. To this end, a preliminary clarification of the concept of learning is needed.

There are various concepts of learning in the economic literature, and their common factor may be expressed in the following way. Let Ω_t be defined as the information-set at time t, and ${}_{t-n}\Omega_t$ as the 'information flow' in the time interval from $t-n$ to t, so that

$$\Omega_{t-n} \cup {}_{t-n}\Omega_t = \Omega_t \tag{4}$$

For the sake of simplicity, it can be assumed – as is usual in economic models – that no loss of information is possible (for a memory failure or a breakdown in the systems of information storage, and so on). Therefore

$$\Omega_{t-n} \subseteq \Omega_t \tag{5}$$

In other words, the stock of information – the information-set – cannot shrink over time. On the basis of these premises, there has been effective learning in the relevant period whenever

$$\Omega_{t-n} \subset \Omega_t \tag{6}$$

that is, the information-set at the beginning of the period turns out to be a proper subset of the information-set at the end of the period.

The simplest concept of learning that complies with the general definition (6) is the updating of the information-set. This is achieved by adding the most recent values of the relevant deterministic variables – or realizations of the relevant stochastic variables – to the information-set:

$$\Omega_{t-1} \cup {}_{t-1}\Omega_t = \Omega_t, \qquad {}_{t-1}\Omega_t \neq \emptyset \tag{7}$$

We define 'real-time updating' as the instantaneous updating that is performed without involving any delay between the new events and their registration in the information-set. This simplistic and demanding concept of learning is the only one consistent with substantive rationality. The other concepts of learning, including non-instantaneous updating, involve systematic errors that are excluded by definition under the assumptions of substantive rationality. A delay in updating would induce even fully rational agents to formulate mistaken expectations and therefore to make suboptimal choices. This eventuality is excluded by the assumption of unbounded rationality, which is intrinsic to substantive rationality. Unbounded rationality

implies that agents have a complete set of relevant information that is constantly updated in real time. Real-time updating is thus not only consistent with substantive rationality, but also a necessary condition of it. However, this condition is merely assumed without justification. To provide behavioural foundations for real-time updating, we should consider the full dynamics of the Interaction, without restricting the analysis to its equilibrium values. Generally speaking, the new events shift the equilibrium of the process. Therefore, a delay in the updating of the information-set would induce a pragmatic error that would lead to lower than expected wealth. Thus, there is an economic incentive to update the information-set immediately by eliminating the cognitive error and the ensuing pragmatic error. The value v of updating from $t-1$ to t, ${}_{t-1}\Omega_t$, is given by

$$v({}_{t-1}\Omega_t) = \max u(x_i \mid \Omega_t) - \max u(x_i \mid \Omega_{t-1}) \tag{8}$$

where $x_i \varepsilon X$ is an option belonging to the set of options X, and $\max u(x_i|\Omega_t)$ is the utility deriving from the best choice within the available option-set X, given the information-set Ω_t. In other words, the value of updating is equal to the increment of utility made possible by the additional information (Willinger, 1989). This value is in general positive, unless the new information is redundant, as in the case of the mean values of stationary variables. This implies the paradox that real-time updating is a necessary condition of substantive rationality, but that such updating has no value, since it applies legitimately only to stationary processes (section 4). Therefore, substantive rationality remains without economic foundations within its own assumptions.

The assumptions underlying procedural rationality allow an analysis of genuine learning – that is, of the mechanism of correcting systematic errors – while the assumptions of behavioural rationality allow an analysis of other aspects of human subjectivity, such as intuition, framing and motivations. Much experimental work has been pursued from this perspective, confirming that economic agents do not typically comply with the tenets of substantive rationality. However, so far, no general theory of learning has emerged from this research. In this chapter, we limit ourselves to showing that the economic role of genuine learning can be analysed only by assuming a concept of rationality more general than that of substantive rationality.

Learning may have a strategic economic value because it permits the exploitation of new information so that a more profitable strategy can be substituted for an existing strategy.[12] Let us assume that ${}_tS_{t+h}$ is a contingent strategy chosen at time t within the time horizon $t+h$. It is possible that, in the light of the new, larger, information-set induced by strategic learning, at time $t+n$ a new strategy is discovered with an expected value $v({}_{t+n}s_{t+h} \mid \Omega_{t+n})$ that exceeds that of the old strategy, chosen at time t, and recalculated in the light of the new information-set: $v({}_ts_{t+h} \mid \Omega_{t+n})$. Therefore, the value of

strategic learning from t to $t+n$, defined as ${}_tV_{t+n}$ as assessed at time $t+n$, may be defined as the difference between the value of the optimal strategy at time $t+n$, in the light of (i) the new enlarged information-set, and (ii) the value of the optimal strategy chosen at time t and reassessed at time $t+n$:

$$ {}_tV_{t+n} = v({}_{t+n}s_{t+h} \mid \Omega_{t+n}) - v({}_ts_{t+h} \mid \Omega_{t+n}) \geq 0 \tag{9} $$

Typically this value is not negative, because the updated information-set may offer new opportunities that were non-existent or unclear beforehand. However, to calculate the net value of strategic learning, it is necessary to account for the costs c_l associated with learning, such as the cost of the acquisition of new information. Therefore, the net value of strategic learning V', neglecting the subscripts for simplicity, may be defined in the following way:

$$ V' = V - c_l \tag{10} $$

A positive net value for strategic learning in an uncertain and open world is a sufficient economic motivation for implementing it. However, to justify a change in strategy, we must also account for the transition costs c_i associated with it, such as transaction costs. The new optimal strategy will typically be implemented only when

$$ V' > c_i \tag{11} $$

Expressions (10) and (11) are distinct because, by assumption, the outcome of strategic learning has been defined as a permanent acquisition, whereas the transition costs are contingent. Therefore, when (11) is not currently satisfied, we cannot exclude the possibility that a fall in the transition costs will justify a change of strategy.

Strategic learning implies the possibility of systematic errors *ex post*, not necessarily *ex ante*, whenever the existing information is efficiently utilized. In the absence of systematic errors *ex post*, learning would be deprived of any strategic value and would become meaningless, at least from the economic point of view. This is the case of perfect foresight and rational expectations.

Thus far, we have considered strategic learning from the perspective of procedural rationality, as adaptation to a world that is possibly characterized by the emergence of new states that do not depend on the conscious design of agents. Learning is thus a genuine addition of relevant knowledge on the World and its evolution. The set of options, however, is not modified by the conscious will of the decision-maker, who is assumed to be an option-taker. From the viewpoint of behavioural rationality, the creativity of *homo sapiens*, which is lacking in *homo economicus*, becomes crucial. Many biologists contend that *homo sapiens* is the only living species that adapts to the environment both passively and proactively; that is by modifying the

environment according to a design. In this view, therefore, economic agents create new options, in that they are able to add new elements to the option-set, based on a design or modification of the World. The value of strategic learning is thus potentially higher because it permits the discovery of a better strategy in the light not only of a growing knowledge of the World and its evolution, but also of the increasing extension of the option-set X_t:

$$ {}_tV'_{t+n} = v({}_{t+n}s_{t+h} \mid X_{t+n} \cap \Omega_{t+n}) - v({}_ts_{t+h} \mid X_t \cap \Omega_{t+n}) - c_i \tag{12} $$

This relationship suggests that behavioural rationality implies a strategic value of learning that is in principle higher than in the case of procedural rationality. In fact, in the case of behavioural rationality, *homo sapiens* learns not only to choose the best option within a given option set to adapt better to a changing environment, as in the case of procedural rationality, but also to extend the set of options for improving the environment itself. This reflects and motivates the increasing capacity of economic agents to be innovative in the technological, organizational and institutional processes in which they are involved.

6 Ontological assumptions and decision theory under uncertainty

The subjects, in our case economic agents, define themselves and their goals in relation to the perceived features of the World. Therefore, the assumptions formulated with respect to its relevant characteristics are crucial for determining the behaviour of economic agents and should be accurately analysed by the economists who observe, interpret and forecast it. We will now examine these neglected aspects of economic methodology with respect to decision theory.

The ultimate foundations of standard economic theory, which grow out of methodological individualism, rely on decision theory. This is self-evident in microeconomics, which is often nothing but decision theory applied to the specific economic problems of individuals or, by extension, families or firms. The analysis of strategic interaction between single decision-makers is based on game theory. However, such theory is basically an application of decision theory to this field. As for macroeconomics, one of its branches refers to a representative agent whose behaviour is explained and predicted on the basis of microeconomic theory, and thus ultimately of standard decision theory. Another branch emphasizes the importance of disaggregation and the crucial role of the foundations of general equilibrium theory. However, general equilibrium models, in turn, need microeconomic foundations. These have been suggested in terms of game theory and therefore, ultimately, of decision theory.[13]

Mainstream economic theory may claim solid foundations in standard decision theory. The analysis of these foundations is illuminating, because the axioms of standard decision theory clarify the conditions under which mainstream macroeconomic theory may be considered well founded. Thus we are able to assess its empirical scope and limitations.

There are two basic types of standard decision theory: the objectivist theory, suggested by Morgenstern and Von Neumann (1944); and the subjectivist theory – often called Bayesian – suggested by Savage (1954), who built on previous work by De Finetti (1937).[14] Since these theories are axiomatized, their theoretical and empirical scope can be studied rigorously. The two theories are apparently very different, since the first is based on a frequentist concept of probability, and the second is based on a personalist (or subjectivist) concept. As for their scope, textbooks typically claim that the objectivist theory applies when the probabilities are 'known', as in roulette, while the Bayesian theory applies when they are 'unknown', as in horse racing. However, the axioms and ontological implications of the two theories are almost identical (Vercelli, 1999). Both theories refer to a World that is familiar to decision-makers, in the sense that the optimal adaptation to it has already occurred (Lucas, 1986). In addition, such a World is *closed*, in the sense that the decision-makers know the complete array of its possible states and of the possible options, and they know the exact consequences of each choice for each possible state of the World. These assumptions make sense only if the World is stationary and time is unimportant, so that genuine innovations and unexpected structural change are excluded. In both cases, the crucial axiom – called the 'axiom of independence' in the Morgenstern–von Neumann case, and the 'sure-thing principle' in Bayesian theory – ensures intertemporal coherence and absence of systematic errors, guaranteeing the substantive rationality of decision-makers over time. Both theories may admit only a very weak kind of uncertainty, which may be called 'soft uncertainty', such that the beliefs of decision-makers may be expressed by a unique, fully reliable, additive probability distribution. Under these assumptions, it is natural to adopt a decision criterion based on the maximization of expected utility or subjective expected utility.

The more general 'behavioural' concepts of rationality and learning outlined above are, in fact, inconsistent with the tenets of standard decision theory. However, alternative decision theories have been proposed in recent times to explain non-standard behaviour in an open and non-stationary world. These theories assume more general measures of uncertainty such as non-additive probabilities (Schmeidler, 1982; Gilboa, 1987), multiple probabilities (Ellsberg, 1961; Gärdenfors and Sahlin, 1982; Gilboa and Schmeidler, 1989) and fuzzy measures (Ponsard, 1986).[15] These theories may be called theories of decision under 'hard uncertainty', to stress the fact that the beliefs of decision-makers may be represented only by a non-additive probability distribution or by a plurality of additive distributions, none of which may be

considered as fully reliable. The non-additivity of the probability distribution reflects uncertainty aversion; that is the awareness that in an open and non-stationary world, relevant unforeseen and unforeseeable contingencies may occur.

Both standard decision theories have a well-developed intertemporal version, but even in these versions, time is substantially unimportant (Kreps, 1988: 190). In the first period, decision-makers choose an optimal intertemporal strategy that is contingent on future states of the world. To comply with the axioms of the theory that assume intertemporal coherence, this strategy cannot be revised in the subsequent periods (Epstein and Le Breton, 1993). Therefore, these theories can be applied only to a closed and stationary world. By definition, the standard approach is unable to take account of the influence that a certain choice may have on the future 'states of nature' (which is forbidden by Savage's definition of states of nature);[16] on future uncertainty (which would imply the analysis of 'endogenous uncertainty', not considered in the standard approach); and on future-choice sets (to analyse intertemporal flexibility preference, see Kreps, 1988).

The theories of decision-making under uncertainty that were developed by Morgenstern–von Neumann and Savage thus cannot be applied to a genuine process of learning, since many of their crucial assumptions and conclusions become implausible in this context. In particular, the existence of systematic errors is inconsistent with the axiom of independence, or the 'sure thing' in Bayesian theory. In addition, to violate the 'sure-thing principle' and the 'compound-lottery axiom', each of which is necessary for a rigorous use of the expected utility approach (Machina and Schmeidler, 1992: 748, 756; Segal, 1987: 177), it is sufficient to assume non-instantaneous learning.

The intertemporal analysis of decisions under uncertainty must account for a further dimension that is of the utmost importance for economic analysis: the degree of irreversibility that characterizes the consequences of economic decisions. Economic irreversibility is measured by the costs of reversing the consequences of a decision; transaction costs are an ubiquitous example of this. It is irreversibility that makes uncertainty such an important issue in many fields of economics. Uncertainty implies unavoidable *ex post* errors. If these errors were easily remedied – promptly and at a low cost – the value of a normative theory of decision under uncertainty would be quite limited. Irreversibility implies that the consequences of an error have a much higher expected value that must account for the possible reversion costs. Unfortunately, while irreversibility greatly increases the practical importance of normative decision theory under uncertainty, it also prevents the use of standard theories. This is because neither objectivist nor subjectivist theories may be satisfactorily applied to irreversible events. It is generally agreed that objectivist decision theories apply only to stationary processes with stable frequencies. In Bayesian theory, the apparently different requirement of 'exchangeability' implies stationarity. Economic processes are

rarely stationary, since they are often characterized by irreversible structural change. In addition, whenever decision-makers believe that an economic system might be non-stationary, their behaviour becomes non-stationary, even assuming that the exogenous environment is stationary (Kurz, 1994). For the same reason, in the standard theories of decision under uncertainty, the value of strategic learning – involving genuine or 'creative' learning – is zero. The value of strategic learning is zero whenever complete irreversibility is postulated; the postulate of independence that implies dynamic coherence (Epstein and Le Breton, 1993) also implies strict irreversibility of the contingent strategy chosen in the initial period.

7 The normative implications of semantic congruence

We are now in a position to summarize the results obtained so far. We must distinguish several key concepts – rationality, learning, expectations, uncertainty, time reversibility and crucial ontological assumptions – that may be semantically congruent or incongruent. We have argued that there is strict semantic correspondence between the different notions of these various concepts, according to a gradient of ontological and epistemic complexity. These correspondences are summarized in a synoptic table, Table 3.1. Each column lists the notions of the concept indicated at the top, increasing in complexity from top to bottom. The notions of the different concepts listed in the same row are characterized by semantic congruence, while notions on different rows are mutually incongruent. Semantic congruence is a neglected, but important, requisite of a sound theory or model. Two concepts are semantically congruent when they can be applied to the same empirical evidence within the same theoretical and/or modelling framework. The normative strength of semantic congruence does not refer to logic: incongruence, unlike logical inconsistency, implies semantic incompatibilities rather than logical contradictions, although it may be a source of the latter.

Table 3.1 Synoptic table of correspondences

Rationality	*Learning*	*Expectations*	*Uncertainty*	*Time*	*World*
Substantive	Updating	Perfect foresight	Certainty	Reversible	Deterministic
Substantive	Updating	Rational expectations	Soft	Reversible	Closed
Procedural	Adaptive learning	Unconstrained adaptive expectations	Hard	Irreversible	Open
Behavioural	Creative learning	Proactive expectations	Hard	Irreversible	Open

There are two kinds of semantic incongruence. The first kind is between a crucial connotation of a theoretical construct – concept, model or theory – and a different connotation of a certain empirical field that forbids a sound application of the incongruent construct to it. For example, the REH implies stationarity of the relevant stochastic processes and cannot be applied safely to empirical evidence that does not exhibit this property. The second kind of semantic incongruence arises between concepts with different semantic scope, such that they cannot be integrated into the same theory or the same model; this is because they cannot soundly be applied to the same data set. For example, rational expectations and genuine learning cannot be part of the same theory or model, because the first concept implies stationarity while the second implies non-stationarity. Semantic incongruence is a crucial source of misguided application of theory to the empirical evidence. The careful consideration of the requisite of semantic congruence is crucial for defining the empirical scope of concepts, theories and models. In what follows, we utilize the information on semantic congruence that is summarized in the synoptic table to discuss the empirical scope of the various approaches characterizing economics.

As we go down the table, the degree of generality increases with the degree of complexity. The table shows that behavioural rationality is the most comprehensive concept of rationality that may account for the complexity of the real world. Behavioural rationality is applicable in cases of hard uncertainty as well as in the limiting case of soft uncertainty, whatever the degree of irreversibility. Behavioural rationality therefore encompasses procedural rationality as a special case – when the environment is taken as given – and substantive rationality under the further restrictive assumption of equilibrium. Whenever a decision problem is characterized by full reversibility or irreversibility, and uncertainty is soft, it is possible to rely on substantive rationality (the Morgenstern–von Neumann theory for 'roulette-wheel' problems, or the Bayesian theory for 'horse-race' problems). However, the use of these theories is justified only within the broader framework of behavioural rationality, structural learning, hard uncertainty and time irreversibility.

The correspondences summarized in the synoptic table are methodologically compelling. For example, genuine structural learning implies a certain degree of irreversibility and non-stationarity, so that it cannot be analysed within the traditional decision theories under (soft) uncertainty. This must be kept well in mind because, reading down in the table, the complexity of the objects analysed increases, making the application of rigorous methods and sophisticated formal language more and more difficult. Economists need to resist the temptation to which they succumb too often – of applying to complex phenomena powerful formal approaches that are fit only for simple phenomena. To consistently study the complex phenomena that characterize the behaviour of modern economies, specific methods must be developed, diligently and patiently. A case in point in the analytical field is

recent progress in refining decision theory under hard uncertainty. This offers new opportunities for a rigorous analysis of structural learning and creative rationality in an open and evolutionary environment. Another promising example is the development of simulation methods that allow a precise, formal analysis of complex economic behaviour, without requiring a high degree of simplification. Yet another stream of literature is based on experimental methods that attempt to reconstruct effective economic behaviour without overly unrealistic normative constraints.

8 Concluding remarks

In this chapter, we have introduced a sharp distinction between ontological complexity, with respect to the properties of the economic system, and epistemic complexity, with respect to the formal properties of the model that represents it. On the basis of this distinction, we have emphasized the limitations of the standard economic approach, which succeeds in simplifying the complexity of economic models only at the cost of restricting their theoretical and empirical scope. If we want to face the problems raised by the irreducible complexity of the real world, we need to introduce an adequate level of epistemic complexity in our concepts and models. Sources of epistemic complexity are, for example, the non-linearity of a dynamic system, or the non-additivity of a probability distribution. Conversely, there are ontological sources of complexity that prevent simplification. To keep in touch with those examples of epistemic complexity, structural change can be cited as a source of non-linearity in dynamic systems, while non-stationarity and openness of the World can be cited as sources of both. In particular, evolutionary phenomena, which are by definition non-stationary and open processes, and are often characterized by essential non-linearities, imply irreducible complexity.

Complexity has been discussed in this essay mainly from a semantic point of view, that is from the perspective of the correspondence and interaction between the degree of complexity of the objective system to be described or forecasted, and its subjective representation through models or expectations. This approach has clear links with the usual formal measures of complexity mentioned in the introduction to this chapter. In particular, by focusing on the dynamic definition of complexity, it may be inferred that standard economics applies – under further limiting conditions – only if the degree of complexity does not exceed Level 2. In other words, from a dynamic viewpoint, standard economics may be applied only to stable linear systems that have a unique equilibrium (Level 1), or to dynamic systems that have a regular periodic attractor, such as a limit cycle (Level 2). To account for higher levels of complexity, more sophisticated concepts of rationality and learning are required, such as those discussed above.

Epistemic complexity is not a virtue, but a necessity. In our theoretical constructs and models, we must introduce the minimal level of epistemic complexity necessary to take account of the ontological complexity essential for understanding and controlling empirical phenomena. Common to all definitions of complexity are properties that cannot be defined in a simpler context. If these properties are considered essential for analysis, they cannot be ignored. If we insist on ignoring them, as often happens in economic analysis, we risk making systematic errors that may lead us astray. We therefore need to introduce into economic models no less – and possibly no more – than the minimal degree of complexity required by the object. This establishes compelling correspondence between the degree of ontological complexity ascertained in the object, and the degree of epistemic complexity of the concepts utilized in its analysis.

The advice of a famous scientist is relevant here: 'Make things as simple as possible – but no simpler.'[17]

Notes

1 As is well-known, according to Simon (1976: 130) 'behaviour is substantively rational when it is appropriate to the achievement of given goals within the limits imposed by given conditions and constructs'. In standard economics, the goal is conceived as maximization of the objective function; i.e., of utility or profit. As Simon himself emphasized, 'as long as these assumptions went unchallenged there was no reason why an economist should acquaint himself with the psychological literature on human cognitive processes or human choice … [and] the irrelevance of psychology to economics was complete' (*ibid.*: 131).

2 The boundary between the two strategies is not always clear and is often more a matter of language than substance. A small change in a single theoretical assumption may be enough to evoke a completely different conceptual world as has occurred in some recent contributions to decision theory under hard uncertainty. On the contrary, what appears to be a radical change may turn out to be of secondary importance for the empirical scope of an application, as in the case of the distinction between objective-probability and subjective-probability decision theory (see section 6 for a brief discussion of both examples).

3 Simon suggested an alternative approach based on the concept of procedural rationality; the approach was defined in the following way: 'behaviour is procedurally rational when it is the outcome of appropriate deliberation … [and] depends on the process that generated it' (Simon, 1976: 131).

4 This feed-back partially overlaps with what philosophers call the 'hermeneutic circle' (e.g., Gadamer, 1960) the solution of which is highly controversial and divides the analytic and continental schools of philosophical thought.

5 As Arrow (1987: 210) maintains 'rational expectations theory is a stochastic form of perfect foresight'.

6 Also some of the most active supporters of substantive rationality pointed out some of its limitations. In particular, Lucas (1986) recognized that the rational expectations hypothesis applies only to stationary processes. However they are generally sceptical about the possibility of applying what they call the 'scientific method' to more complex situations (see Vercelli, 1991).

7 For this reason, we may refer to this more general concept of behavioural rationality as 'designing rationality' (Vercelli, 1991).
8 Some advocates of procedural rationality give a broader definition that encompasses what we call here 'behavioural rationality'.
9 This difficulty was already fairly clear in Samuelson (1947). Scarf (1960) provided precise examples of plausible trading processes that fail to exhibit global stability. Sonnenschein (1973) proved that under the usual assumptions about consumer preferences and behaviour, not even quasi-global stability can be assured. This result was confirmed and extended by subsequent papers. A brief survey of this literature may be found in Mas-Colell, Whinston and Green (1995).
10 This equation may be generalized by assuming that the variables represented are vectors of variables.
11 A comprehensive survey of this growing stream of literature may be found in Evans and Honkapohja (2001).
12 This definition of strategic learning intends to 'capture' its economic motivations and not to exclude further, and perhaps more important, motivations.
13 Even proponents of the branches of macroeconomics that do not accept methodological individualism have often tried to provide foundations in terms of various types of decision theory based on assumptions that falsify the assertions of standard theory and justify alternative theoretical assertions. A case in point is Shackle (1952), who elaborated a non-standard decision theory as the foundation of his heterodox interpretation of Keynesian economics.
14 Anscombe and Aumann (1963) proposed a sort of synthesis of the two basic varieties. The discussion of the latter applies also to this stream of literature.
15 For surveys relevant to the argument developed here, see, for example, Kelsey and Quiggin (1992) and Vercelli (1995, 1999).
16 According to Savage, the states of nature are independent of the actions of agents.
17 This statement was uttered by Albert Einstein, according to Phillips, Freeman and Wicks (2003: 486).

References

Albin, P. (2001) *Barriers and Bounds to Rationality: Essays on Economic Complexity and Dynamics in Interactive Systems* (Princeton, NJ: Princeton University Press).
Anscombe, F.J. and R.J. Aumann (1963) 'A Definition of Subjective Probability', *Annals of Mathematical Statistics*, vol. 34, pp. 199–205.
Arrow, K.J. (1953) 'Le rôle des valeurs boursières pour la répartition la meilleure des risques', *Econométrie*, vol. 11, pp. 41–8.
—— (1987) 'Rationality of Self and Others in an Economic System', in R.M. Hogart and M.W. Reder (eds), *Rational Choice. The Contrast between Economics and Psychology* (Chicago: University of Chicago Press), pp. 201–16.
Arrow, K.J. and G. Debreu (1954) 'Existence of an Equilibrium for a Competitive Economy', *Econometrica*, vol. 26, pp. 265–90.
Arrow, K.J. and F.H. Hahn (1971) *Competitive General Equilibrium Analysis* (San Francisco: Holden-Day).
Begg, D.K.H. (1982) *The Rational Expectations Revolution in Macroeconomics* (Oxford: Allan).
Bray, M. (1982) 'Learning, Estimation, and the Stability of Rational Expectations Equilibria', *Journal of Economic Theory*, vol. 26, pp. 313–17.
Bray, M. and N.E. Savin (1986) 'Rational Expectations Equilibria, Learning and Model Specification', *Econometrica*, vol. 54, pp. 1129–60.

Cagan, P. (1956) 'The Monetary Dynamics of Hyperinflation', in M. Friedman (ed.), *Studies in the Quantity Theory of Money* (Chicago: Aldine), pp. 25–117.
DeCanio, S.J. (1979) 'Rational Expectations and Learning from Experience', *Quarterly Journal of Economics*, vol. 93, pp. 47–58.
Debreu, G. (1959) *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, Cowles Foundation Monograph no. 17 (New Haven, Conn.: Yale University Press).
De Finetti, B. (1937) 'La Prévision: Ses Lois Logiques, Ses Sources Subjectives', *Annales de l'Institut Henri Poincaré*, vol. 7, pp. 1–68. English translation: 'Foresight: Its Logical Laws, Its Subjective Sources', in H.E. Kyburg and H.E. Smokler (eds) (1964) *Studies in Subjective Probabilities* (New York: Wiley), pp. 93–158.
Ellsberg, D. (1961) 'Risk, Ambiguity, and the Savage Axioms', *Quarterly Journal of Economics*, vol. 75, pp. 643–69.
Epstein, L. and M. Le Breton (1993) 'Dynamically Consistent Beliefs Must Be Bayesian', *Journal of Economic Theory*, vol. 61, pp. 1–22.
Evans, G.W. (1983) 'The Stability of Rational Expectations in Macroeconomic Models', in R. Frydman and E.S. Phelps (eds), *Individual Forecasting and Aggregate Outcomes. 'Rational' Expectations Examined* (Cambridge: Cambridge University Press), pp. 67–94.
Evans, G.W. and S. Honkapohja (1994) 'Learning, Convergence and Stability with Multiple Rational Expectations Equilibria', *European Economic Review*, vol. 38, pp. 1071–98.
——(2001) *Learning and Expectations in Macroeconomics* (Princeton, NJ: Princeton University Press).
Foley, D.K. (2001) 'Introduction', in Albin (2001), pp. 3–72.
Fourgeaud, C., C. Gourieroux and J. Pradel (1986) 'Learning Procedure and Convergence to Rationality', *Econometrica*, vol. 54, pp. 845–68.
Gadamer, H.-G. (1960) *Wahrheit und Methode* (Tübingen: J.C.B. Mohr) English translation: *Truth and Method* (London and New York: Sheed & Ward, 1989).
Gärdenfors P. and N.-E. Sahlin (1982) 'Unreliable Probabilities, Risk Taking, and Decision Making', *Synthèse*, vol. 53, pp. 361–86.
Gilboa, I. (1987) 'Expected Utility with Purely Subjective Non-Additive Probabilities', *Journal of Mathematical Economics*, vol. 16, pp. 65–8.
——(1989) 'Additivizations of Nonadditive Measures', *Mathematics of Operation Research*, vol. 4, pp. 1–17.
Gilboa, I. and D. Schmeidler (1989) 'Maximin Expected Utility with a Non-unique Prior', *Journal of Mathematical Economics*, vol. 18, pp. 141–53.
——(1993) 'Updating Ambiguous Beliefs', *Journal of Economic Theory*, vol. 59, pp. 33–49.
Ingrao, B. and G. Israel (1990) *The Invisible Hand: Economic Equilibrium in the History of Science* (Cambridge, Mass.: MIT Press).
Kahneman, D. (2003) 'Maps of Bounded Rationality: Psychology for Behavioral Economics', *American Economic Review*, vol. 93, pp. 1449–75.
Kelsey D. and J. Quiggin (1992) 'Theories of Choice under Ignorance and Uncertainty', *Journal of Economic Surveys*, vol. 6(2), pp. 133–53.
Keynes, J.M. (1936) *The General Theory of Employment, Interest and Money* (London: Macmillan, now Palgrave) and as Vol. 7: *The Collected Writings of John Maynard Keynes* (London: Macmillan, now Palgrave, 1973).
Kreps, D.M. (1988) *Notes on the Theory of Choice* (Boulder, CO: Westview Press).
Kurz, M. (1994a) 'On Rational Belief Equilibria', *Economic Theory*, vol. 4, pp. 859–76.
——(1994b) 'On the Structure and Diversity of Rational Beliefs', *Economic Theory*, vol. 4, pp. 877–900.

Kurz, M. (1995) 'Rational Preferences and Rational Beliefs', in K.J. Arrow, E. Colombatto, M. Perlman and C. Schmidt (eds), *The Rational Foundations of Economic Behaviour* (London: Palgrave Macmillan), pp. 339–61.
Lucas, R.E. Jr. (1978) 'Asset Prices in an Exchange Economy', *Econometrica*, vol. 46, pp. 1429–45.
—— (1986) 'Adaptive Behavior and Economic Theory', *Journal of Business*, vol. 59, Supplement, pp. 5401–26.
Machina, M.J. and D. Schmeidler (1992) 'A More Robust Definition of Subjective Probability', *Econometrica*, vol. 60(4), pp. 745–80.
Marcet, A. and T.J. Sargent (1988) 'The Fate of Systems with "Adaptive" Expectations', *American Economic Review, Papers and Proceedings*, vol. 78, pp. 168–72.
—— (1989a) 'Convergence of Least Squares Learning Mechanisms in Self-Referential Stochastic Models', *Journal of Economic Theory*, vol. 48, pp. 337–68.
—— (1989b) 'Convergence of Least Squares Learning in Environments with Hidden State Variables and Private Information', *Journal of Political Economy*, vol. 97, pp. 1306–22.
March, J.G. (1988) *Decisioni e organizzazioni* (Bologna: Il Mulino).
Mas-Colell, A., M.D. Whinston and J.R. Green (1995) *Microeconomic Theory* (Oxford: Oxford University Press).
Minford, P. and D. Peel (1983) *Rational Expectations and the New Macroeconomics* (Oxford: Martin Robertson).
Morgenstern, O. and J. von Neumann (1944) *The Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press).
Muth, J.F. (1961) 'Rational Expectations and the Theory of Price Movements', *Econometrica*, vol. 29, pp. 315–35.
Phillips, R., R.E. Freeman and A.C. Wicks (2003) 'What Stakeholder Theory is Not', *Business Ethics Quarterly*, vol. 13(4), pp. 479–502.
Ponsard, C. (1986) 'Foundations of Soft Decision Theory', in J. Kacprzyk and R.R. Yeger (eds), *Management Decision Support Systems Using Fuzzy Sets and Possibility Theory* (Cologne: Verlag TUV).
Samuelson, P.A. (1947) *Foundations of Economic Analysis* (Cambridge, Mass.: Harvard University Press).
Sargent, T.J. (1993) *Bounded Rationality in Macroeconomics* (Boston, MA: MIT Press)
Savage, L.J. (1954) *The Foundations of Statistics* (New York: John Wiley & Sons), Revised and enlarged edition (New York: Dover, 1972).
Scarf, H. (1960) 'Some Examples of Global Instability of Competitive Equilibrium', *International Economic Review*, vol. 1(3), pp. 157–72.
Schmeidler, D. (1982) 'Subjective Probability Without Additivity', Working Paper, Foerder Institute for Economic Research, Tel Aviv University.
Schumpeter, J.A. (1934) *The Theory of Economic Development* (Oxford: Oxford University Press).
Segal, U. (1987) 'The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach', *International Economic Review*, vol. 28(1), pp. 175–202.
Shackle, G.L.S. (1952) *Expectations in Economics* (Cambridge: Cambridge University Press).
Simon, H.A. (1976) 'From Substantive to Procedural Rationality', in S.J. Latsis (ed.), *Method and Appraisal in Economics* (Cambridge: Cambridge University Press), pp. 129–48; reprinted in Simon, H.A. (1982) *Models of Bounded Rationality*, vol. 2, pp. 424–43.
—— (1982) *Models of Bounded Rationality* (Cambridge MA: MIT Press).

Sonnenschein, H. (1973) 'Do Walras' Identity and Continuity Characterize the Class of Community Excess Demand Functions?', *Journal of Economic Theory*, vol. 6, pp. 345–54.

Vercelli, A. (1991) *Methodological Foundations of Macroeconomics. Keynes and Lucas* (Cambridge: Cambridge University Press).

——(1995) 'From Soft Uncertainty to Hard Environmental Uncertainty', *Economie Appliquée*, vol. 48, pp. 251–69.

——(1999) 'The Recent Advances in Decision Theory under Uncertainty: A Non-technical Introduction', in L. Luini (ed.), *Uncertain Decisions: Bridging Theory and Experiments* (Dordrecht: Kluwer), pp. 237–60.

——(2000) 'Financial Fragility and Cyclical Fluctuations', *Structural Change and Economic Dynamics*, vol. 1, pp. 139–56.

——(2002) 'Uncertainty, Rationality and Learning: A Keynesian Perspective', in S. Dow and J.Hillard (eds), *Keynes, Uncertainty and the Global Economy*, Vol. 2 (Cheltenham: Edward Elgar), pp. 88–105.

Walras, L. (1874) Elément d'économie politique pure ou théorie de la richesse sociale (Lausanne: L. Corbaz) trans. W. Jaffé, *Elements of Pure Economics or The Theory of Social Wealth* (London: Allen & Unwin), 1954.

Willinger, M. (1990) 'Irréversibilité et cohérence dynamique des choix', *Revue d'Economie Politique*, vol. 100(6), pp. 808–32.

Woodford, M. (1990) 'Learning to Believe in Sunspots', *Econometrica*, vol. 58, pp. 277–307.

4

Altruism: Evolution and a Repercussion

Oded Stark, You Qiang Wang and Yong Wang

Under the theme: 'Altruism, Evolution and a Repercussion' we take the unusual step of including two papers, one by Oded Stark and You Qiang Wang, the other by Oded Stark and Yong Wang. The papers represent two interrelated aspects of a unified research project on the evolution and consequences of altruism. The first part explains the formation of the altruistic trait and the second part focuses on its implications. Accordingly, the first paper we have included here 'On the Evolutionary Edge of Altruism: A Game-Theoretic Proof of Hamilton's Rule for a Simple Case of Siblings' relates to the first part of the project and seeks to explain the prevalence of altruism. The second paper, 'The Intergenerational Overlap and Human Capital Formation', relates to the second part of the project. This paper already assumes that individuals are altruistically inclined and explores the repercussion of altruism within families on the formation of human capital.

Editors, Bina Agarwal and Alessandro Vercelli

Part 1 On the evolutionary edge of altruism: a game-theoretic proof of Hamilton's rule for a simple case of siblings*

Oded Stark
Universities of Bonn, Klagenfurt and Vienna; Warsaw University; Social Research Center, Cologne, Germany and Eisenstadt, Austria
and
You Qiang Wang
Tsinghua University, China

* Reprinted with the permission of Springer Verlag, from *the Journal of Evolutionary Economics*, vol. 14, pp. 37–42 (February 2004). We are indebted to an anonymous referee and to Uwe Cantner for helpful comments and suggestions. Partial financial support from the National Institute on Aging (grant RO1-AG13037) and from the Humboldt Foundation is gratefully acknowledged.

1 Introduction

Evolutionary biologists have developed a powerful theory of the evolutionary foundations of altruism between relatives. The theory is based on the idea that individuals who are related by blood share genes. Consider a gene that governs a particular behaviour. The likelihood that the gene will be replicated is higher when the gene takes into account not only the extra reproductive opportunities that the behaviour confers on the host who carries the gene, but also the extra reproductive opportunities that the behaviour confers on relatives of the host who also carry the gene. William Hamilton, the pioneer of this theory, describes it as follows:

> The social behaviour of a species evolves in such a way that in each distinct behavior-evoking situation the individual will seem to value his neighbor's fitness against his own according to the coefficients of relationship appropriate to that situation. (Hamilton, 1964: 19)

The coefficient of relationship between two individuals is the probability that a randomly selected gene in one of these individuals will have an exact copy located in the other individual as a result of descent from a common ancestor. In the case of a haploid population in which each parent has a single gene for being altruistic or selfish and mating is monogamous, the coefficient of relationship between two siblings is $\frac{1}{2}$. 'Hamilton's rule' is that altruism will spread in a population if the benefit obtained from giving multiplied by the coefficient of relationship exceeds the cost of giving. If c is the cost to oneself of helping a sibling, and b is the benefit to a sibling from receiving help, altruism will spread if $b \cdot \frac{1}{2} > c$, that is, if the benefit obtained from help exceeds twice the cost of helping.

The purpose of this part of the chapter is to complement the large and important literature that followed Hamilton's pioneering articles, both in evolutionary biology (notably Dawkins, 1976; Grafen, 1984; and Wilson, 1987) and beyond (Axelrod, 1984; Hofbauer and Sigmund, 1988; Binmore and Samuelson, 1992; and Nowak and May, 1992), with a formal game-theoretic proof of Hamilton's rule. Building on Bergstrom and Stark (1993), Bergstrom (1995) and Stark (1999), this part provides a proof of the rule for a simple case of siblings.

In evolutionary economics, the study of altruism is motivated by two questions: where does altruism come from and what does it give rise to? The incentive to explore inclinations is not independent, however, from the density of implications. If the motivation to produce, the propensity to redistribute and the tendency to accumulate and transfer – within families, societies and across generations – matter both for individual well-being and for social welfare, and if these processes are governed or significantly affected by the incidence and intensity of altruism as a trait, we would like

to find out how the trait evolves. The interest in economics, and beyond, in the evolution, survival and extinction of institutions of various types cannot be orthogonal to the interest in the prevalence and intensity of altruism *if* altruism gives rise to patterns and predispositions that completely or partially substitute for institutional mandates, impinge on the design of institutions, crowd out their roles, or render their mission superfluous. All the more so when there is a close correspondence between altruism and cooperation.[1] Since altruism is practiced and manifested socially, it is natural to start the search for its prevalence and origins in small social groupings such as the family. It is more likely that altruism will pervade large groupings such as the population at large if it evolves between siblings than if it fails to gain a foothold even within families.

2 The game and a general result

In each period there is an old generation and a young generation. A fraction of the old generation consists of altruists, a complementary fraction consists of nonaltruists. Members of the old generation are matched with uniform probabilities into pairs. Each pair breeds two children. The children constitute the young generation. The two siblings play a one-shot prisoner's dilemma game with each other. A sibling can help the other sibling at a cost to himself. Let c be the cost to a sibling of helping a sibling, and let b be the benefit to the sibling who receives the help, $b > c > 0$. We obtain the following payoff matrix:

		Column sibling	
		C	D
Row sibling	C	$b-c, b-c$	$-c, b$
	D	$b, -c$	$0, 0$

where playing C stands for providing help and playing D stands for not offering help. A sibling who plays C is altruistic, a sibling who plays D is nonaltruistic. To see this suppose the column sibling selects C. If the row sibling selects C rather than D, he give up b to receive the smaller $b-c$, whereas the column sibling gains since he receives $b-c$ which is larger than $-c$. Suppose, alternatively, that the column sibling selects D. Again, if the row sibling selects C rather than D, his payoff declines (by c), while the column sibling's payoff rises (by b). This defines altruism: giving up something for the sake of another. Thus, throughout the rest of this part of the chapter we identify altruism with playing cooperatively in the one-shot prisoner's dilemma game.

Let $(p, 1-p)$ denote the mixed strategy in which the row sibling plays C with probability p; and let $(q, 1-q)$ denote the mixed strategy in which the column sibling plays C with probability q. Then, for any given (p, q), the expected payoffs of the row and column siblings are $qb - pc$ and $pb - qc$, respectively. Let p_1 and p_0 be the probabilities that the row sibling plays C if the column sibling plays C and D, respectively. We now provide a game-theoretic proof of Hamilton's rule for a simple case of siblings.

Proposition (Hamilton's rule). If c is the cost to oneself of helping a sibling, and b is the benefit to a sibling from receiving help, altruism will spread if $b \cdot (p_1 - p_0) > c$, that is, $p_1 - p_0$ is the equivalent of the coefficient of relationship.

Proof Cooperation will be globally stable if the expected payoff of a randomly selected cooperator child is larger than the expected payoff of a randomly selected defector child. Without loss of generality, we randomly select the column sibling. The expected payoff of a randomly selected cooperator column sibling is $p_1 b - c$ since $(p, q) = (p_1, 1)$, and the expected payoff of a randomly selected defector column sibling is $p_0 b$ since $(p, q) = (p_0, 0)$. Hence, cooperation will be globally stable if $p_1 b - c > p_0 b$, yielding the statement of the proposition.

3 The rule of imitation, survival and the outcome: a special case

We assume that how a child plays, C or D, is determined through the imitation of his parents, and that each child imitates one of his parents with equal probabilities. The probability that a child survives to reproduce (to have his own children) is proportional to the payoff in the game. For example, consider a case in which the payoff positively influences the probability of reaching maturity and of being able to procreate.

Let x be the proportion of cooperative parents, and let $1 - x$ be the proportion of defector parents.

Claim 1 The probability that a randomly chosen child is a cooperator is x.

Proof Let N be the number of individuals in the old generation. Hence, the number of parent pairs is $N/2$. Cooperator children come either from cooperator–cooperator parent pairs or from cooperator–defector parent pairs. The number of cooperator–cooperator marriages is $(N/2)x^2$. All $(N/2)x^2 \cdot 2 = Nx^2$ children of these marriages are cooperators. The number of cooperator–defector marriages is $(N/2)[x(1-x) + (1-x)x] = Nx(1-x)$ which is also the number of cooperator children produced by such marriages. Hence, the total number of cooperator children is

$Nx^2 + Nx(1 - x) = Nx$. In a population of N children, the probability then that a randomly chosen child is a cooperator is $Nx/N = x$.

Claim 2 Given that a child is a cooperator, the conditional probability that its sibling is a defector is $(1 - x)/2$.

Proof A cooperator–defector pair of children results from a mixed marriage. Half of these $Nx(1 - x)$ marriages produce mixed sibling pairs. The number of cooperating children in the mixed sibling pairs from these marriages is $\frac{1}{2}Nx(1 - x)$. As already shown, the total number of cooperator children is Nx. Given that a child is a cooperator, the conditional probability that its sibling is a defector is $\frac{1}{2}Nx(1 - x)/Nx = (1 - x)/2$.

Given that a child is a cooperator, the conditional probability that its sibling is a cooperator is the complementary probability $1 - (1 - x)/2 = (1 + x)/2$.

By replacing x for $1 - x$ we get that given that a child is a defector, the conditional probability that its sibling is a cooperator is $x/2$. (Given that a child is a defector, the conditional probability that its sibling is a defector is the complementary probability $1 - (x/2)$.)

Claim 3 Cooperation will be globally stable if $b > 2c$.

Proof Since in this case $p_1 = (1 + x)/2$ and $p_0 = x/2$, the claim follows from the proposition.

Evolutionary biologists refer to 'inclusive fitness' of an individual. In the present model the inclusive fitness of a cooperator parent is the expected payoff of a randomly selected cooperator child, and the inclusive fitness of a defector parent is the expected payoff of a randomly selected defector child. The inclusive fitness of a cooperator is larger than the inclusive fitness of a defector, and cooperation is globally stable if the benefit to a child from playing cooperatively (helping) exceeds twice the child's own cost of playing cooperatively (helping).

4 Conclusion

We have shown that in a simple case of siblings, Hamilton's rule can be derived as the outcome of a prisoner's dilemma game between siblings. We employed several simplifying assumptions. These may be relaxed. For example, the formation of couples can be more selective than random. As shown in the Appendix, however, this change will only strengthen the case for cooperation.

Part 2 The intergenerational overlap and human capital formation*

Oded Stark
Universities of Bonn, Klagenfurt and Vienna; Warsaw University; Social Research Center, Cologne, Germany and Eisenstadt, Austria
and
Yong Wang
City University of Hong Kong, China

1 The idea

It is well-recognized that the stock of human capital affects the level of per capita output in an economy. Whether the effect arises because human capital is an ordinary input in the economy's production function or because the effect manifests itself through the enhancement of total factor productivity (in that it leads to the creation, adoption, implementation and diffusion of new technologies) are largely empirical issues. The notion that an economy that forms a large quantity of human capital will have a higher per capita output than an economy that forms a small quantity of human capital can safely be taken as given, requiring little, if any, additional inquiry. But why is it that one economy has, or forms, abundant per capita human capital, while another has, or forms, little? Why does the per capita human capital gap between economies not close? Much – though not all – of the human capital in an economy is the result of decisions made by individuals. Clearly, several factors are involved and one of them is life expectancy: a longer life expectancy entails a longer payback period that, in turn, encourages larger investments in human capital. An economy consisting of individuals with a long life expectancy will then form more human capital than an economy consisting of individuals with a short lifespan.

The impact of a lengthened life expectancy comes from the returns side of the human capital investment calculus: the marginal benefit is higher. We argue however that, typically, a lowered marginal cost of forming human capital is imbedded in a lengthened life expectancy. We seek to unearth this effect and study its role in accounting for the divergent experiences of economies in the formation of human capital. We suggest that the lowered marginal cost effect arises from a correlate of extended life expectancy: the prolonged duration of the overlap between generations. Suppose that as long as they are alive, parents support the human capital formation of their children, and that the parental support is cheaper than market financing.[2] An extended life expectancy that results in a prolonged overlap entails more parental support,

* Partial financial support from the Humboldt Foundation and the Sohmen Foundation is gratefully acknowledged.

which in turn can foster the formation of more human capital. An example will serve to illustrate.

Suppose that life expectancy is 45. An individual gives birth to one child when the individual is 20 years of age. The child is cared for in his infancy and for as long as he engages in acquiring human capital, conditional on the individual being alive. The age at which the child makes the human capital formation decision is 15. At this age, if the child were to engage in human capital formation, the child could expect parental support for up to 10 years. If the child finds it optimal to devote more than 10 years to human capital formation, he can do so by borrowing at a fixed market interest rate. When the child reaches the age of 20, he gives birth to a child whom he, in turn, will support in the same manner in which he was supported. Suppose that the child finds it optimal to acquire human capital for a little more than 10 years, say for τ years in excess of 10. During these years the child has to bear the entire cost of forming human capital, which includes the market rate of interest.

Suppose now that life expectancy is 55. Retaining all other assumptions as before, the child can now expect parental support for up to 20 years. To see the implications of this assumption for human capital formation, consider the case $0 < \tau < 10$. All of the years of human capital formation previously financed by commercial loans now become parentally supported, interest-free years. Since the marginal cost of forming human capital goes down, more human capital will be formed. This effect is separate from the *returns* to human capital, a marginal benefit that arises from the addition of years during which returns to the human capital investment can be reaped.[3]

In section 2 we present our analytical framework. In section 3 we formally investigate the effect of extended overlapping on the formation of human capital by optimizing individuals. To this end we decompose the 'gross' life-expectancy effect into a 'net' life-expectancy effect and an overlapping effect. In section 4 we trace the welfare implication of extended overlapping for an economy that is subjected to such a change. In section 5 we further explain the rationale underlying our idea and offer a suggestion as to how to differentiate empirically between the overlapping model of human capital formation and the received model of human capital formation.

2 The analytical framework

Consider a continuous overlapping-generations economy akin to that of Cass and Yaari (1967). At every instant of time a generation is born. A generation consists of a continuum of individuals of measure N. The lifespan of an individual is l. The individual gives birth to a child after time spell l^c $(0 < l^c < l)$ has elapsed. Thus, each member of generation t has a single parent in generation $t - l^c$, and a single offspring in generation $t + l^c$. At each point in time the economy consists, therefore, of a continuum of overlapping generations,

each at an age between 0 and l. The economy's population size is thus a constant of measure Nl.

During their lifetime, individuals form human capital, work and procreate. Let an individual spend a portion of his lifetime immediately following birth forming human capital and the complementary portion of his lifetime working. While the acquisition of human capital is costly, as it entails the opportunity cost of forgone wage earnings, it subsequently enhances the individual's productivity, and hence his earnings. Since an individual gives birth to a child after l^c (≥ 0) of his lifespan has elapsed, $l^p \equiv l - l^c$ measures the duration of the overlap between the individual and his child.[4,5]

Let s_t represent the time span that an individual of generation t chooses to allocate to human capital formation. Hence the remaining $l - s_t$ of the individual's lifespan is allocated to work. Let the cost of forming human capital be a proportion λ of the individual's wage. The cost of forming human capital is born by the individual's parent as long as the parent is alive, and by the individual himself through borrowing at the market interest rate if additional human capital is formed past the parent's death.[6] When the individual reaches the age l^c, he has a child of his own. That child too is faced with a choice of allocating his lifetime between human capital formation and work, drawing on his parent's support in a manner akin to that described earlier, that is, up to a duration of l^p. The amount of human capital (measured in efficiency units of labour) that is available to the individual and supplied by him inelastically, generated by the allocation of time s_t to human capital formation, is given by $\varphi(s_t)$ where $\varphi(0) = 1; \varphi(s_t) > 1, \varphi'(s_t) > 0, \varphi''(s_t) < 0$ for $s_t \in (0, l)$; $\lim_{s_t \to 0} \varphi'(s_t) = \infty$, and $\lim_{s_t \to l} \varphi'(s_t) = 0$. The assumption $\varphi(0) = 1$ is made in order to incorporate the feature that the individual is endowed with one efficiency unit of labour (the individual's pair of hands) that is available to the individual even if no human capital is formed.

Let r_t and w_t be the instantaneous interest rate and the instantaneous wage rate at time t, respectively. The lifetime income (in present-value terms) of a generation t individual who chooses to invest s time in human capital formation, recalling the method of financing described above, is

$$V_t = \int_{t+s}^{t+l} d_t^\tau \varphi(s) w_\tau d\tau - \int_{t+\min(s,l^p)}^{t+s} d_t^\tau \lambda w_\tau d\tau - \int_{t+l^c}^{t+l^c+\min(s',l^p)} d_t^\tau \lambda w_\tau d\tau$$

where $d_t^\tau \equiv e^{-\int_t^\tau r_\upsilon d\upsilon}$ is the discount factor at time t for wages received and costs incurred at future time τ, and s' is the duration of the human capital formation period, chosen by the individual's child. Without loss of generality,[7] we assume that the individual seeks to maximize his lifetime income, that is,

$$s_t = \arg\max_s V_t \tag{1}$$

Hence, the optimal human capital formation span for an individual of generation t is implicitly given by the first-order condition

$$\varphi'(s_t)\int_{t+s_t}^{t+l} d_t^{\tau} w_\tau d\tau = d_t^{t+s_t} w_{t+s_t}\left[\varphi(s_t) + \lambda\delta(s_t - l^p)\right] \tag{2}$$

where $\delta(x) = 0$ for $x < 0$, $\delta(x) = 1$ for $x > 0$, and $0 \leq \delta(0) \leq 1$.

We now briefly describe the economy. We have in mind a small economy that operates in a perfectly competitive world in which economic activity extends over an infinite continuous time. At every point in time the economy produces a single consumption good using perfectly durable (physical) capital and labour measured in efficiency units in the production process. The supply of funds for investment purposes at each point of time consists of domestic savings and of net international borrowing. Funds are supplied by, and can be borrowed from, a perfectly competitive world capital market at the stationary positive rate of return to capital, $\bar{r}$, in terms of the consumption good. The supply of labour at each point of time is the sum of the aggregate supply of human-capital-augmented labour of all the generations. Production at each point of time occurs according to a constant-returns-to-scale production function which is invariant across time. Therefore, the output produced at time t, Y_t, is

$$Y_t = F(K_t, L_t) \equiv L_t f(k_t); \qquad k_t = K_t/L_t \tag{3}$$

where K_t and $L_t = \int_{\tau+s_\tau \leq t} \varphi(s_\tau) N d\tau$ are the capital and labour employed at time t, respectively. The production function $f(k)$ is strictly increasing and strictly concave. Producers operate in a perfectly competitive environment. Profit maximization gives rise to the following first-order conditions

$$r_t = f'(k_t) \tag{4}$$

$$w_t = f(k_t) - f'(k_t)k_t \tag{5}$$

where r_t and w_t are the interest rate and wage rate at time t, respectively, and output is the numeraire. Given the unrestricted nature of the international capital markets, the economy's interest rate is exogenously given at the world level $\bar{r}$, at all times. Consequently, the capital–labour ratio employed in production is stationary at a level $\bar{k}$,[8]

$$\bar{k} = f'^{-1}(\bar{r}) \tag{6}$$

and the wage rate is stationary at a level $\bar{w}$,

$$\bar{w} = f(\bar{k}) - f'(\bar{k})\bar{k} \tag{7}$$

Since the economic environment in which the individual optimizes is stationary, with $d_t^\tau = e^{-\bar{r}(\tau - t)}$, we can ignore time subscripts and rewrite (2) as

$$\frac{1}{\bar{r}}[1 - e^{-\bar{r}(l-s)}]\varphi'(s) = \varphi(s) + \lambda\delta(s - l^p) \tag{8}$$

The left-hand side of (8), multiplied by $\bar{w}$, measures the marginal benefit of human capital formation. The right-hand side of (8), multiplied by $\bar{w}$, measures the marginal cost of human capital formation, which has two components. The first component is the usual opportunity cost of forgone earnings and the second component reflects the impact of the overlap between parents and children. When the duration of the period of human capital formation chosen by the individual is shorter than the duration of the overlap, $\delta(s - l^p) = 0$; the entire cost of human capital formation is borne by the individual's parent and the second component vanishes. However, when the individual chooses to form human capital for a time span that is longer than the duration of the overlap with his parent, $\delta(s - l^p) = 1$; the marginal cost of forming human capital incorporates the extra cost of financing human capital formation through the marketplace. Given the assumptions concerning the production function of human capital, the solution (8), and the solution to the individual's maximization problem in (1), is unique and interior (that is, the length of time allocated to human capital formation maintains $s \in (0, l)$).[9]

3 The effect of an extended intergenerational overlap on human capital formation

Although in this chapter we are interested in investigating the consequences of the duration of the overlap between parents and children as measured by l^p, typically an increase in l^p arises from the prolongation of life expectancy l. Therefore, a change in l affects the endogenous variables through two channels: changing the life expectancy, and varying the duration of the overlap between parents and children. We are able, though, to separate the effects of a change in l on the investment in human capital that arises from these two channels. While there are many interesting models that focus on the link between human capital formation and life expectancy (recent examples include Stark, 1999, chapter 2; Kalemli-Ozcan, Ryder and Weil, 2000; Leung and Wang, 2003), our investigation of the overlapping-duration channel is novel.

Suppose that there is an increase in l. The left-hand side of (8), as a function of s, shifts upward: the conventional life-expectancy channel is at work. As a result of the increase in l, individuals live longer and hence are able to reap the returns to human capital formation over a longer period, raising the marginal benefit of human capital formation. But the increase in l also increases l^p by

the same amount (keeping l^c constant), which in turn affects the marginal cost of human capital in the right-hand side of (8) in a more subtle way. The right-hand side of (8), as a function of s, is a smooth function except for a vertical jump that occurs at $s = l^p$. Upon an increase in l^p, the jump in the marginal cost function occurs at a later point in time, thereby extending the range within which the marginal cost of human capital formation is low. This is due to the overlapping-duration channel: a larger l^p implies a longer overlap between parents and children, which in turn allows children to enjoy parental support for forming human capital for a longer period of time. To the extent that the extra parental support lowers the cost of forming human capital at the margin, the overlapping-duration channel is operative as it encourages additional human capital formation that would not have been possible had the intergenerational overlap remained the same. In short, while the life-expectancy channel operates from the benefit side, the overlapping-duration channel operates from the cost side.

Suppose that l rises from l_1 to l_2, and hence l^p rises from l_1^p to l_2^p. The following figures illustrate circumstances in which the overlapping-duration channel is fully operative.

Figure 4.1 shows that individuals initially choose $s_1^* < l_1^p$ for engaging in human capital formation so that their entire investment is paid for by the parents and no market finance takes place. Following the increase of l from l_1 to l_2, the marginal benefit curve shifts up due to the life-expectancy effect, and the marginal cost curve extends the range (the darkened segment between l_1^p and l_2^p) within which human capital formation is family-financed (the overlapping-duration effect). Consequently, individuals choose s_2^*. Had the cost structure of human capital formation been the same as before (that is, without the overlapping-duration effect), the life-expectancy effect alone

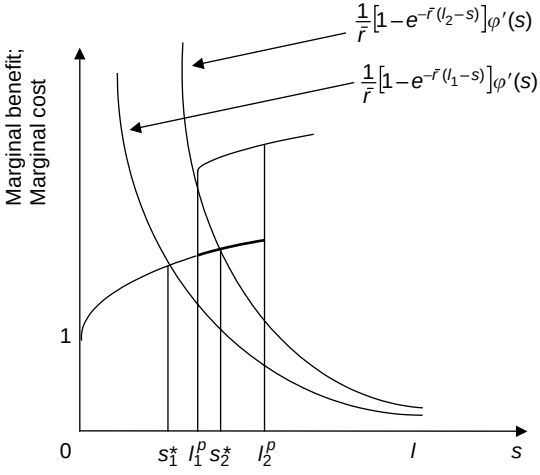


Figure 4.1 The life-expectancy effect and the overlapping-duration effect: Case I

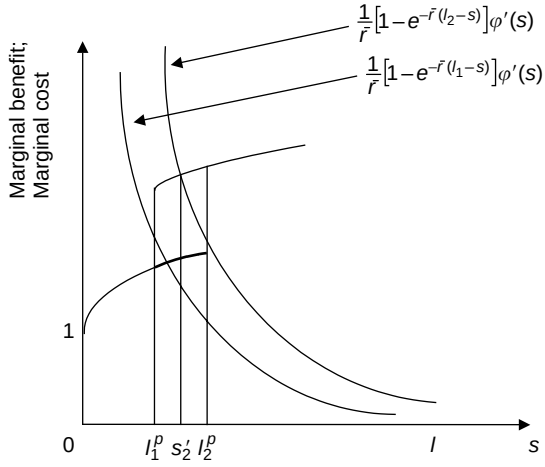


Figure 4.2 The life-expectancy effect and the overlapping-duration effect: Case II

would have resulted in a duration of human capital formation equal to l_1^p, which is less than s_2^*. Hence the additional period of human capital formation of $s_2^* - l_1^p$ can be attributed to the pure effect of the overlapping duration.

In Figure 4.2, individuals initially choose $s_1^* = l_1^p$ for human capital formation so that the constraint of parentally-supported human capital formation just binds. Following an increase of l from l_1 to l_2, and of l^p from l_1^p to l_2^p, individuals choose l_2^p. In this case the life-expectancy effect results in a duration of human capital formation that is equal to only s_2', and the overlapping-duration effect contributes to the additional increase of $l_2^p - s_2'$ in the duration of human capital formation.

Similarly, Figure 4.3 illustrates the case in which individuals initially choose $s_1^* > l_1^p$, relying on market financing above and beyond the overlapping period with their parents. Following the increases in l and l^p, they choose l_2^p, wherein the overlapping-duration effect again contributes to the additional increase of $l_2^p - s_2'$ in the duration of human capital formation.

Having provided a non-exhaustive list of cases in which the overlapping-duration channel is operative in human capital formation decisions, we should add that, of course, the overlapping-duration channel is not always operative. Nonetheless, the combined effect of the life-expectancy channel and the overlapping-duration channel is always positive. This can be stated as the following proposition.

Proposition 1 An increase in l will always lead to an increase in human capital formation, that is, $(\partial s/\partial l) > 0$.

Proof See Appendix.

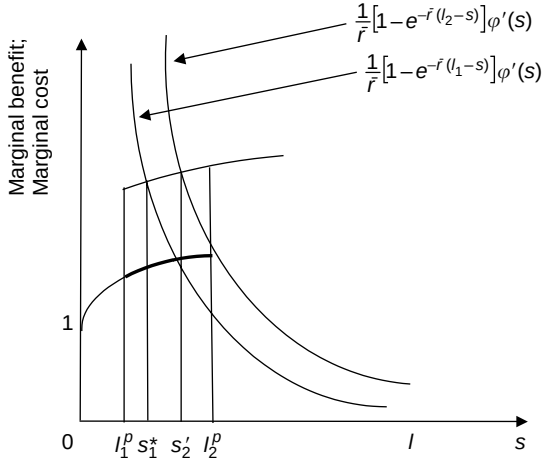


Figure 4.3 The life-expectancy effect and the overlapping-duration effect: Case III

Our argument so far amounts to a statement that the positive effect of a longer life expectancy on human capital formation arises from two distinct effects: a pure life-expectancy effect and a prolonged intergenerational overlapping effect. Yet decomposing an effect into its constituent parts falls short of demonstrating that each part has a life of its own. Thus, we next investigate the pure overlapping-duration effect, studying its role in isolation from, and independently of, the conventional life-expectancy effect.

Suppose that individuals give birth to their children at a somewhat earlier age while their life expectancy remains intact. This change entails an increase in l^p that is not associated with a change in l. While, by construction, the left-hand side of (8) remains unaltered so that the life-expectancy channel is not operative, the change in l^p affects the right-hand side of (8) through the overlapping-duration channel. To illustrate the pure effect of the overlapping duration on human capital formation, suppose that l^p increases from l_1^p to l_2^p (keeping l constant). To facilitate comparison we consider once again three cases wherein the initial choice of the duration of the human capital formation span is less than, equal to, or greater than the duration of overlap.

In Figure 4.4, individuals initially choose to form human capital for a period that is shorter than the duration of the overlap with their parents, $s_1^* < l_1^p$. In this case a prolonged overlapping has no impact on the individuals' decision as to how much human capital to form.

In Figure 4.5, the initial decision is to set the period of human capital formation equal to the duration of the overlap, that is, $s_1^* = l_1^p$. In this case, the marginal benefit curve intersects the vertical portion of the marginal cost curve, and the extended overlap has a clear and positive effect – it increases the individuals' human capital formation period to s_2^*.[10]

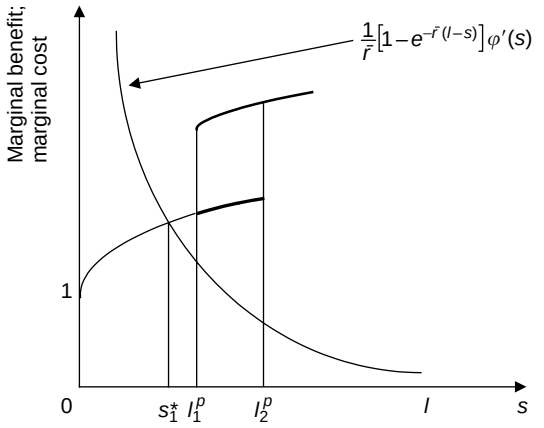


Figure 4.4 The pure overlapping-duration effect: Case I

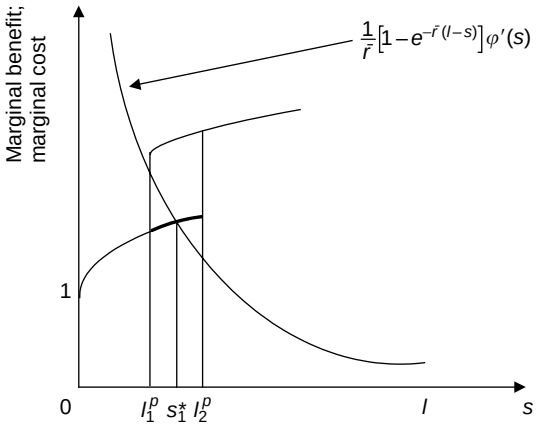


Figure 4.5 The pure overlapping-duration effect: Case II

Lastly, Figure 4.6 presents the case where individuals initially choose a duration of human capital formation, s_1^*, that exceeds the duration of the overlap with their parents. The extended overlap prompts additional human capital formation, provided that the increase in the duration of the overlap is large enough (that is, as large as $l_2^p > s_1^*$). We summarize these results on the pure effect of overlapping in the following proposition.

Proposition 2 An increase in l^p from l_1^p to l_2^p without any change in l leads to a strict increase in human capital formation by an individual, that is, $s_2^* > s_1^*$, if $s_1^* = l_1^p$ or if $l_1^p < s_1^* < l_2^p$.

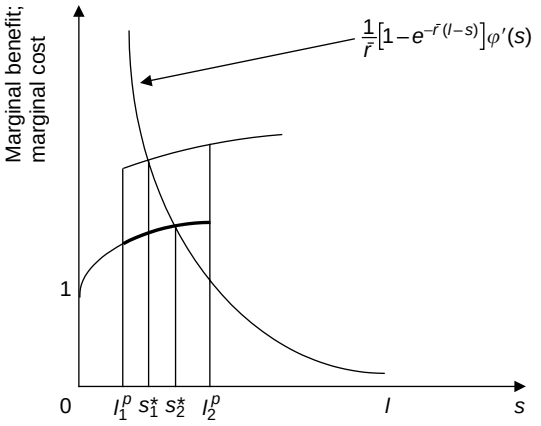


Figure 4.6 The pure overlapping-duration effect: Case III

4 The welfare effect of an extended intergenerational overlap

We have shown that the duration of the intergenerational overlap of individuals with their parents can impact positively on an individual's formation of human capital. This channel of influence is independent from the usual repercussion of the life-expectancy channel.

To analyse the welfare effect of an extended overlap, both when it operates in conjunction with the life-expectancy effect and when it operates independently of the life-expectancy effect, we first calculate the flow of per capita output. Since at any given point in time the economy's labour input measured in efficiency units is $L = \int_{t-l}^{t-s} \varphi(s)Nd\tau = (l-s)\varphi(s)N$, the economy-wide output (given (6)) is $Y = Lf(\bar{k})$, and population size is Nl, then per capita output is

$$y = \frac{Y}{Nl} = \frac{(l-s)\varphi(s)f(\bar{k})}{l} \tag{9}$$

Measuring welfare by the flow of per capita output in (9), we present our results regarding the welfare implication of an extended intergenerational overlap in the two cases, that is with and without a simultaneous change in life expectancy, in the following proposition:

Proposition 3 (i) For a given $\Delta l = \Delta l^p > 0$, $\Delta y > 0$ holds; and (ii) for a given $\Delta l^p > 0$ but $\Delta l = 0$, $\Delta y > 0$ holds as long as $\Delta s > 0$. In both cases, Δy is larger the larger is Δs.

Proof (i) It suffices to show that $\Delta y > 0$ for any small increase of Δl in both l and l^p. Total differentiation of (9) yields

$$\Delta y = \frac{[(\Delta l - \Delta s)\varphi(s) + (l-s)\varphi'(s)\Delta s]l - (l-s)\varphi(s)\Delta l}{l^2} f(\bar{k})$$
$$= \frac{f(\bar{k})}{l}\left([(l-s)\varphi'(s) - \varphi(s)]\Delta s + \frac{s\varphi(s)\Delta l}{l}\right)$$

Since $\int_s^l 1 \cdot dt > \int_s^l e^{-\bar{r}(t-s)}dt$, or $l - s > \frac{1}{\bar{r}}[1 - e^{-\bar{r}(l-s)}]$ (assuming a positive interest rate $\bar{r}$), it follows from (8) that $(l-s)\varphi'(s) - \varphi(s) > 0$. Since $\Delta s > 0$ from Proposition 1, it follows that $\Delta y > 0$.

(ii) Since l is constant in this case, denoting $s \equiv s(l^p)$ and differentiating both sides of (9) with respect to l^p, we have

$$\frac{dy}{dl^p} = \frac{f(\bar{k})}{l}[(l-s)\varphi'(s) - \varphi(s)]\, s'(l^p)$$

Hence, for a given $\Delta l^p > 0$, we obtain

$$\Delta y = \frac{f(\bar{k})}{l}\int_{l^p}^{l^p+\Delta l^p} [(l-s)\varphi'(s) - \varphi(s)]\, s'(l^p)dl^p$$
$$= \frac{f(\bar{k})}{l}\int_s^{s+\Delta s} [(l-s)\varphi'(s) - \varphi(s)]\, ds$$

Again, since it follows from (8) that $(l-s)\varphi'(s) - \varphi(s) > 0$, it is clear that $\Delta y > 0$ as long as $\Delta s > 0$.

In both cases, it is easy to see from the expressions of Δy that the larger the increase in s that arises from a given increase in l or l^p, the larger the increase in y.

When an extended overlap, resulting from prolonged life expectancy, brings about additional human capital formation as illustrated in Figures 4.1, 4.2 and 4.3, it also raises the per capita output at any point in time. Similarly, when an extended overlapping that is not accompanied by a change in life expectancy induces additional human capital formation as illustrated in Figures 4.5 and 4.6, it also raises the per capita output at any point in time. Therefore, whenever the overlapping-duration channel is operative, either in conjunction with the life-expectancy effect or independently of it, an increase in the intergenerational overlap is welfare-improving. By the same token, a shortening of the overlap between parents and their children can have an adverse impact on human capital formation, and hence on welfare.

Corollary Consider two identical economies in which the overlapping-duration channel is operative. The economy that experiences an increase

in the duration of the overlap will enjoy a higher per capita output than the economy in which the duration of the overlap remains unchanged.

5 Complementary reflections

In a highly stylized economy in which material capital is the only production input, the production function is concave and the cost of acquiring capital is linear, a lengthening of the lifespan of capital prompts the optimal acquisition of more capital. The pioneers some four decades ago of the modern theory of human capital, notably Jacob Mincer, Theodore W. Schultz and Gary S. Becker, were duly aware of the powerful analogy between the effect of the lifespan of material capital and the effect of longevity, as a proxy of the length of the period during which human capital renders a return. Yet, while the acquisition of more material capital (machinery) today in response to a lengthened lifespan would presumably crowd out the acquisition of material capital tomorrow, a lengthened life expectancy could crowd in human capital formation by the next generation. Here, the direct analogy between the two types of capital apparently breaks down.

The positive effect on human capital formation of overlapping with a parent arises from the parent's provision of support for the child's formation of human capital. In the absence of any reverse transfer from the child to the parent, the motive for the parental support is altruism. In this part of the chapter we *assume* parental altruism within dynasties rather than explain why and how it evolves – an issue that we address in the first part of this chapter, as well as in related work (Falk and Stark, 2001).[11,12]

While the assumption that the parent provides somewhat less than full support in the child's pursuit of human capital will affect the absolute size of the effects in our model, it will not change the model's qualitative predictions.

A widely held view maintains that in developing economies, delayed marriage and postponed childbearing will hasten the pace of economic development and entail a higher per capita output. The rationale is that as a consequence of delay and postponement, the denominator in the output per capita ratio will be smaller, and the numerator will be larger since adults (young women) will be spending more of their productive time working in the economy rather than tending to home production (rearing children). Yet if the economic environment in which such changes occur is characterized by a fixed (or little-changed) life expectancy, the intergenerational overlap will be reduced, possibly impinging *negatively* on human capital formation (Proposition 2), and on welfare (Proposition 3).

To discriminate between the received model of human capital formation and the overlapping model, consider a setting in which the life expectancy of the individual's parent is rising (the intergenerational overlap is lengthened) and the individual's life expectancy is declining, yet the individual invests *more* in human capital formation. Such an outcome can arise only from

the operation of the overlapping effect since the individual's negative life-expectancy effect (the shortened duration of the payback period to an investment in human capital) implies a reduced investment in human capital. The same discriminating test applies if the life expectancy of the individual's parent is rising, and the individual's life expectancy remains unchanged.

One possibility for empirically assessing the distinct effect of the intergenerational overlap on human capital formation is to examine the age of home-leaving and to explore whether this age correlates positively with schooling. A study of the long-term trend in the age of home-leaving in the United States (Gutmann, Pullum-Piñon and Pullum, 2001) provides illuminating evidence. The study asks whether young people aged 15–29 were living with one or both of their parents at the time of each of the decennial censuses from 1880 through 1990. The study finds that, with the exception of the Second World War era (and in contrast to widely-held views), the age of leaving home *rose* in twentieth-century America. The study further suggests that a reason for leaving home early is the death of the parents, and it points out that over the century there was a dramatic decline in the likelihood of becoming an orphan between the ages of 15 to 29, a change brought about by the steady decline in adult mortality. In addition, the study highlights the sharply increased likelihood of attendance of high school by those aged 15–19, 'and with it the likelihood that they would live at home' (p. 10). Thus, the long-term trends that the study depicts are that the age at which young people ceased to live with their parents rose, adult mortality declined, and schooling and higher education – especially in the form of community colleges – increased.

Within the field of the economics of human capital formation, there has long been a debate concerning the causal relationship between education and health (with age being the most important component of health). Many empirical studies have shown that there is a positive correlation between education and health. However, the source of this correlation is not clear. It has been suggested that the observed correlation is caused by a third variable that is correlated both with education and with health (Grossman and Kaestner, 1997; Grossman, 2000). The duration of the intergenerational overlap could constitute the elusive variable.

If the poor in an economy overlap with their children for a shorter time span than do the rich, the children of the poor will run out of parental support earlier than the children of the rich, and could therefore acquire less human capital even if all children have access to equally-priced market finance. Thus, rendering the terms under which children from poor families can borrow in order to pay for their acquisition of human capital equal to the terms under which children from rich families can so borrow may not equalize the investment in human capital environment for the two types of children under differential overlapping.

A low likelihood that a costly human capital formation today will be rewarded by a flow of returns tomorrow dampens investment in human capital.

Among the considerations that impinge on this likelihood is the risk to life emanating from civil strife. It is less appreciated, though, that the probability that civil strife will occur is negatively affected by the level of investment in human capital: people who stand to lose a large quantity of human capital are less inclined to resort to violent means of settling disputes and resolving conflicts than people who risk only meager quantities of human capital. To the extent that an extended overlap entails the formation of a larger quantity of human capital, the duration of the overlap will be correlated negatively with the likelihood of civil strife or with the likelihood of brutality.

Appendix to Part 1

To substantiate the claim that a non-random formation of couples will only strengthen the case for cooperation, note that if matching is purely (positively) assortative, the fractions of cooperator marriages and defector marriages are, respectively, x and $1-x$. To allow matching patterns that are intermediate between the polar cases of purely random matching and purely assortative matching, we define a parameter m where $0 \leq m \leq 1$, such that when matching is purely random $m = 0$, and when matching is purely assortative $m = 1$. The number of cooperator–cooperator marriages is then $(N/2)[x^2 + mx(1-x)]$, and the number of cooperator–defector marriages is $N(1-m)x(1-x)$.

It follows that the probability that a randomly chosen child is a cooperator is x; given that a child is a cooperator, the conditional probability that its sibling is a defector is $(1-m)(1-x)/2$; given that a child is a cooperator, the conditional probability that its sibling is a cooperator is $1-(1-m)(1-x)/2$; and given that a child is a defector, the conditional probability that its sibling is a cooperator is $(1-m)x/2$. (Given that a child is a defector, the conditional probability that its sibling is a defector is $1-(1-m)x/2$.) Since in this case $p_1 = 1-(1-m)(1-x)/2$ and $p_0 = (1-m)x/2$, it follows from the proposition that cooperation will be globally stable if $b[1-(1-m)(1-x)/2-(1-m)x/2] > c$ that is, if $b(1+m) > 2c$.

Appendix to Part 2: proof of Proposition 1

We first consider the case where either $s < l^p$ or $s > l^p$, and then the case where $s = l^p$.

If either $s < l^p$ or $s > l^p$, that is, the marginal benefit curve intersects the marginal cost curve either at its lower portion before the jump (recall Figure 4.1), or at its upper portion after the jump (recall Figure 4.4), the second term in the right-hand side of (8) is constant. Hence, from differentiating both sides of (8) with respect to l we obtain

$$e^{-\bar{r}(l-s)}\left(1-\frac{\partial s}{\partial l}\right)\varphi'(s)+\frac{1}{\bar{r}}[1-e^{-\bar{r}(l-s)}]\varphi''(s)\frac{\partial s}{\partial l}=\varphi'(s)\frac{\partial s}{\partial l}$$

which implies that $\partial s/\partial l = (e^{-\bar{r}(l-s)}/A)\varphi'(s)$, where

$$A=[1+e^{-\bar{r}(l-s)}]\varphi'(s)-\frac{1}{\bar{r}}[1-e^{-\bar{r}(l-s)}]\varphi''(s)>0$$

Thus, $(\partial s/\partial l) > 0$ for both $s < l^p$ and $s > l^p$.

Considering now the case where $s = l^p$, we prove, by contradiction, that $\Delta s > 0$ for a given $\Delta l = \Delta l^p > 0$. Suppose this is not true so that $\Delta s \leq 0$ for a given $\Delta l = \Delta l^p > 0$.

Since $s = l^p$, we obtain from (8) that

$$\frac{1}{\bar{r}}[1 - e^{-\bar{r}(l-s)}]\varphi'(s) = \varphi(s) + \lambda\delta(0)$$

where $\delta(0) \geq 0$. Let $l_1 = l + \Delta l$ $(>l)$, $l_1^p = l^p + \Delta l^p$ $(>l^p)$, and $s_1 = s + \Delta s$ $(\leq s)$. It then follows that $s_1 < l_1^p$ and hence, after the increases in l and l^p, (8) becomes

$$\frac{1}{\bar{r}}[1 - e^{-\bar{r}(l_1-s_1)}]\varphi'(s_1) = \varphi(s_1)$$

But, because $\varphi(\cdot)$ is an increasing and concave function, the following inequality also holds:

$$\frac{1}{\bar{r}}[1 - e^{-\bar{r}(l_1-s_1)}]\varphi'(s_1) > \frac{1}{\bar{r}}[1 - e^{-\bar{r}(l-s)}]\varphi'(s) = \varphi(s) + \lambda\delta(0) \geq \varphi(s_1)$$

The apparent contradiction completes the proof.

Notes

1 The inquiry pursued in this part of the chapter relates to the study of institutions in yet another way. Schelling (1960, 1971, 1978) has shown how the interactions of individuals in environments characterized by bounded rationality and imperfect information coalesce over time into customs, norms and institutions that govern economic and social life. Schelling's pioneering work was recently supplemented significantly by Young's study of economic and social institutions. To Young (1998) an institution is an established law, custom, usage, practice, organization. (Examples of institutions are aplenty: rules of the road, time of lunch, patterns of marriage, forms of economic contracts.) Young develops a theory that predicts how institutions evolve and characterizes their welfare properties. Viewing Hamilton's rule as an institution places this part of the chapter's inquiry in that research vein.

2 We assume that the technology of human capital formation is invariant to the method of financing; the edge of parental support over market financing arises not from parents' direct involvement in the formation of human capital by their children, but from the intergenerational transfer constituting a means of financing human capital acquisition that is less expensive than the market-based means of financing.

3 To a child at the age of 15, the expected overlap with a parent whose age is 35 is best given by the contemporaneous life expectancy of adults at the age of 35 rather than by the life expectancy of the parental generation at birth. Life expectancy at birth is quite sensitive to the incidence of children dying at very early ages. Historically, life expectancy was increased through reductions in the number of children dying during infancy and the sharp increases in life expectancy at birth were not followed by corresponding increases in longevity, although the two measures were positively correlated. Since in this part of the chapter our interest is in the effect of changes in the lifespan of adults, that is, in changes in the mean age of death beyond infancy, our reference to life expectancy in the sections that follow should be understood as life expectancy net of the effect of infant mortality.

4 Alternatively, it can be assumed that the individual gives birth to a child at a younger age and that the child reaches the human capital formation age only at a point in time that is l^c into the individual's life. The years prior to that

point in time are immaterial since they do not affect the child's human capital formation decision.

5 Since our aim is to unravel the pure effect of alternative durations of overlapping on human capital formation, we consider the timing of giving birth, l^c, as exogenously given.

6 Our interest in this part of the chapter is in human capital formation. We concentrate on the effect of *inter vivos* transfers on human capital formation and we exclude bequests. While the bequests that individuals receive undoubtedly affect their welfare, ordinarily bequests are received at a point in time in individuals' life that is long past their human capital formation years.

7 In a more general setting in which the individual maximizes his lifetime utility $\int_t^{t+l} e^{-\beta(\tau-t)}u(c_\tau)d\tau$ subject to the budget constraint $\int_t^{t+l} d_t^\tau c_\tau d\tau \leq V_t$, where $u(\cdot)$ is the instantaneous utility function, c_τ is consumption at time τ, and β is the discount rate for the preferences, it is rather straightforward to show that under a constant interest rate and perfectly competitive markets, assumptions to which we resort subsequently, the individual's indirect utility is strictly increasing in his lifetime income V_t.

8 The assumptions of a small open economy and of perfect capital mobility imply that the individual may borrow and lend at the constant world interest rate of $\bar{r}$. This result conveniently rules out the possibility of a complex dynamics – the economy is always at a steady state. In particular, if there is a shock to the world interest rate $\bar{r}$, the economy will respond by moving to the steady state that is associated with the new interest rate instantaneously.

9 The second-order condition for a maximum holds: $(1/\bar{r})[1 - e^{-\bar{r}(l-s)}]\varphi''(s) - [1 + e^{-\bar{r}(l-s)}]\varphi'(s) < 0$.

10 We exclude from consideration the exceptional case in which the marginal benefit curve intersects the marginal cost curve at the lower corner of its vertical portion. In other words, when $s = l^p$, we assume that s solves (8) with a $\delta(0) > 0$.

11 The extended overlap is tantamount to enhanced altruism, an effect studied in Stark (1999, chapter 1). There, as here, the effect on the child's well-being is positive.

12 We further assume that parental altruism takes the form of sharing a meal, not imposing a will; parents do not decide for their children how much human capital the children should form.

References

Axelrod, R. (1984) *The Evolution of Cooperation* (New York: Basic Books).

Bergstrom, T.C. (1995) 'On the Evolution of Altruistic Ethical Rules for Siblings', *American Economic Review*, vol. 85, pp. 58–81.

Bergstrom, T.C. and O. Stark (1993) 'How Altruism Can Prevail in an Evolutionary Environment', *American Economic Review,* vol. 83, pp. 149–55.

Binmore, K.G. and L. Samuelson (1992) 'Evolutionary Stability in Repeated Games Played by Finite Automata', *Journal of Economic Theory*, vol. 57, pp. 278–305.

Cass, D. and M. Yaari (1967) 'Individual Saving, Aggregate Capital Accumulation, and Efficient Growth', in K. Shell (ed.), *Essays on the Theory of Optimal Growth* (Cambridge, MA: MIT Press), pp. 233–68.

Dawkins, R. (1976) *The Selfish Gene* (Oxford and New York: Oxford University Press).

Falk, I. and O. Stark (2001) 'Dynasties and Destiny: The Roles of Altruism and Impatience in the Evolution of Consumption and Bequests', *Economica*, vol. 68, pp. 505–18.

Grafen, A. (1984) 'Natural Selection, Kin Selection and Group Selection', in J.R. Krebs and N.B. Davis (eds), *Behavioural Ecology: An Evolutionary Approach* (Oxford: Blackwell Scientific), pp. 62–84.

Grossman, M. (2000) 'The Human Capital Model', in A.J. Culyer and J.P. Newhouse (eds), *Handbook of Health Economics* (Amsterdam: North-Holland), pp. 347–408.

Grossman, M. and R. Kaestner (1997) 'Effects of Education on Health', in J.R. Behrman and N. Staccy (eds), *The Social Benefits of Education* (Ann Arbor: University of Michigan Press), pp. 69–123.

Gutmann, M.P., S.M. Pullum-Piñon and T.W. Pullum (2001) 'Three Eras of Young Adult Home Leaving in Twentieth-Century America', mimeo, University of Texas at Austin, Texas.

Hamilton, W.D. (1964) 'The Genetical Evolution of Social Behaviour', Parts I and II, *Journal of Theoretical Biology*, vol. 7, pp. 1–52.

Hofbauer, J. and K. Sigmund (1988) *The Theory of Evolution and Dynamical Systems* (Cambridge: Cambridge University Press).

Kalemli-Ozcan, S., H. Ryder and D. Weil (2000) 'Mortality Decline, Human Capital Investment, and Economic Growth', *Journal of Development Economics*, vol. 62, pp. 1–23.

Leung, C.M.M. and Y. Wang (2003) 'Endogenous Health Care, Life Expectancy, and Economic Growth', mimeo, City University of Hong Kong, China.

Nowak, M.A. and R.M. May (1992) 'Evolutionary Games and Spatial Chaos', *Nature*, vol. 359, pp. 826–9.

Schelling, T.C. (1960) *The Strategy of Conflict* (Cambridge, MA: Harvard University Press).

Schelling, T.C. (1971) 'Dynamic Models of Segregation', *Journal of Mathematical Sociology*, vol. 1, pp. 143–86.

Schelling, T.C. (1978) *Micromotives and Macrobehaviour* (New York: Norton).

Stark, O. (1999) 'Siblings, Strangers, and the Surge of Altruism', *Economics Letters*, vol. 65, pp. 135–42.

Stark, O. (1999) *Altruism and Beyond: An Economic Analysis of Transfers and Exchanges Within Families and Groups* (Cambridge: Cambridge University Press).

Wilson, D.S. (1987) 'Altruism in Mendelian Populations Derived from Sibling Groups: The Haystack Model Revised', *Evolution,* vol. 41, pp. 1059–70.

Young, H.P. (1998) *Individual Strategy and Social Structure*: *An Evolutionary Theory of Institutions* (Princeton, NJ: Princeton University Press).

5

Human Reproduction and Utility Functions: An Evolutionary Approach*

Alexander A. Vasin
Moscow State University, Faculty of Computational Mathematics and Cybernetics, and New Economic School, Russia

1 Introduction

The basic models of game theory and economics involve individual utility or payoff functions. Each player or participant is characterized by his set of strategies and exogenously-given payoff function. He independently sets his strategy, which influences not only his payoff, but also the payoffs of other participants. The models describe an individual's behaviour as aimed at maximizing his payoff function. The theory studies methods and outcomes of rational strategic choices. A standard assumption is that each player knows the payoff functions of all participants. The case of incomplete information about the payoff functions of other players is also studied; for instance, through the Bayesian (see Fudenberg and Tirole, 1991) and maximin (see Germeyer, 1976) approaches. Note that under both complete and incomplete information the payoff functions are exogenously given and do not change.

Conventional assumptions about utility functions continue to be used when game theory is applied to economics or sociology. However, these standard assumptions, which influence research results, do not always accord with reality. In particular, a typical textbook on economics proceeds from the concept of *homo economicus*. As a producer, such a person aims to maximize

* A large part of this chapter was prepared during my visit to the Carlos III University, Madrid, in 2002. I am grateful to the Department of Economics for its hospitality. Financial support of this research by the Ministry of Education, Culture and Sport of Spain and the Russian Fund for Basic Research grant 02-01-00610 is gratefully acknowledged. I thank participants in the seminars at Alicante, Bilbao and Madrid, for useful discussion. My special thanks to Jim Leitzel and the editors of this volume for many useful proposals on refinement and clarification of the text.

his profit, while as a consumer he likes to increase his own consumption. If a model includes the labour market, then the individual utility function also typically includes hours of labour, where more labour is associated with lower utility (for instance, see Myles, 1995).

Recently several papers have studied the Russian labour market within this standard framework (Slinko, 1999; Friebel and Guriev, 2000). The puzzle they addressed was the survival in the 1990s of a large number of state-owned enterprises (such as libraries, and research and cultural organizations) where wages were considerably lower than in the private sector. The researchers identified non-monetary distribution mechanisms as the main reason for low labour mobility. However, the explanatory power of the models turned out to be rather low. The explanation appears to lie in the lack of correspondence between the actual behaviour of a large part of the population and standard assumptions about individual utility functions in labour economics. Some people are willing to work for a smaller wage than they can earn in another occupation because they find their current jobs more interesting or more useful for society. Other individuals are motivated not by the absolute level of income they earn, but by whether they can achieve an income within a desirable range of the overall income distribution in society. For such individuals, the huge income inequality in Russia could suppress their labour mobility. Yet others might be willing to forego higher pay for a job where they can communicate with colleagues whom they might see as substitutes for a family.

There are additional problems with conventional theory. All game-theoretic solutions (including Nash equilibria and other non-cooperative and cooperative principles) proceed from the following paradigm: in a strategic situation an individual should maximize his/her payoff while taking into account the power and interests of other participants. However, if we observe social practice we find that it deviates from this standard view of decision-making. It shows that individual interests are variable and that the most efficient way for a participant to achieve his goals might be to change the payoff functions of other participants. For example aggressive television advertising which makes people desire more and new goods and drugs can alter utility functions.

The present chapter examines whether it is possible to endogenize the determination of payoff functions and to identify the mechanisms that change them. One field that provides the relevant models and ideas is evolutionary game theory. An evolutionary game is a model of behavioural dynamics in some human or animal population. In economic theory, evolutionary games were first employed to justify the concepts of Nash equilibrium and dominance elimination for interactions among bounded rational individuals (see Vasin, 1989, 1991; Nachbar, 1990; and Samuelson and Zanq, 1992). Another purpose was to distinguish between stable and unstable Nash equilibria. These early models included exogenously given utility functions

and did not consider the relation between population behaviour dynamics and reproduction.

Section 2 below focuses on evolutionary games that describe behaviour in self-reproducing populations, where every new individual is a child of some adult, and fertility and mortality rates depend on individual strategies. The behavioural dynamics of the population are determined by demographic values and by the *evolutionary mechanism*. In general, the evolutionary mechanism involves some combination of genetic inheritance, adaptation, imitation and education, determining the strategies of newborns and changes in the strategies of adults. For instance, *replication* means that a child inherits the behaviour strategy of his parent (of the same sex), and an adult does not change his strategy during his lifetime. The main conclusion from replication studies is that 'natural' payoff functions of individuals concern reproduction irrespective of the specific population or the type of interaction. More precisely, an evolutionary stable distribution over strategies includes only those strategies that maximize individual fitness – the sum of fertility and viability, or the rate of reproduction of individuals employing one of the strategies in the stable distribution. (The concept of evolutionary stable behaviour was introduced by Maynard Smith, 1982.) Loosely speaking, an evolutionary stable distribution is such that, under any sufficiently small deviation, the evolutionary process converges back to the distribution.

My result that evolutionary stable distributions consist only of individual fitness-maximizing strategies is consistent with previous results found using Replicator Dynamics. However, the difficulty remains that the endogenous utility function in general depends upon the particular evolutionary mechanism. In order to address this problem I construct a model of natural selection of evolutionary mechanisms themselves. The model shows that if replication takes place among competing evolutionary mechanisms, then an evolutionary stable strategy combination includes only strategies that maximize individual fitness: only evolutionary mechanisms that are compatible with individual fitness maximization survive under competition. Ok and Vega Redondo (2001) obtain a similar result for a model of individual preference evolution. Their model considers a special type of evolutionary mechanism with inheritance of preferences. Different results were obtained in a model where an individual can observe the preference of his partner and choose his strategy accordingly. In this case the evolutionary stable strategy maximizes the *total fitness* (that is, the number of descendants) of the partners. However, the observability assumption is problematic, especially because an individual would have an interest in manipulating the available information concerning his preferences (see Samuelson, 2001, on this issue). Only under special conditions can an individual reliably identify the preferences (or the evolutionary mechanism) of his partner. One such case is interaction among close relatives. I consider a model where an individual distinguishes between his relatives (siblings, cousins, etc.) and strangers, and can choose his strategy

depending on these distinctions. The model studies the evolution of altruistic and cooperative behaviour; it shows that such behaviour among relatives is evolutionary stable if it maximizes the total fitness of the family. I also discuss factors that may limit the spread of such behaviour.

Demographic data show that behaviour in modern social (human) populations maximizes neither individual nor family or population fitness, although these populations are apparently self-reproducing. Section 3 discusses why the evolutionary models do not apply directly to these populations. Perhaps the most important reason is that social behaviour is very complicated, so replicating behavioural strategies is impossible in actual practice. I consider mechanisms that provide the opportunity to influence individual behaviour and discuss who or what can change individual utility functions. I introduce the concept of a *superindividual*: a self-reproducing object that uses a human population as a resource for its own reproduction and can influence the evolution of behaviour within this population. Corporations might be considered one type of superindividual. Several papers consider corporate competition within an evolutionary process (see Vega-Redondo, 1997), but I am not aware of any advanced studies of corporation impact on individual behaviour. I argue that competition among superindividuals is the crucial force for behaviour evolution in modern society, and discuss the role of an individual in a context where he cannot reproduce himself, but can choose the superindividuals he helps to reproduce.

2 Evolutionary models of behaviour in self-reproducing populations

Evolutionary game theory was primarily developed with respect to biological populations (Maynard Smith, 1982). However, its models and approaches are also of interest for social modelling. This section considers several models of behaviour dynamics amongst interacting individuals. The result of interaction in every time period determines birth and death rates for individuals depending on their strategies. In all the models, biological reproduction is the force behind the endogenous determination of individual utility functions.

The first model is the well-known Replicator Dynamics (RD). It assumes that a newborn inherits the strategy of his parent. The results below show that given a long enough time period, only those with strategies that maximize the sum of individual fertility and viability survive in the population. Thus, the surviving individuals act as if this sum ('individual fitness', in the terminology of Charles Darwin) is their utility function, although the model does not assume any rational choice of strategies.

The second model considers a society of interacting individuals where populations differ in their evolutionary mechanisms and hence in the dynamics of the strategy distributions within these populations. Every evolutionary mechanism includes a rule determining the individual strategy choice of

a newborn and a rule governing how the strategies of adults change. These rules may include different types of imitation and rational choices based on utility functions. One interesting conclusion is that if RD is among the competing evolutionary mechanisms, then it determines the evolution of behaviour: individual fitness turns out to be an endogenous utility function since after a long enough time period, in any surviving population individuals act as if they maximize their fitness.

The second model also assumes that individuals cannot observe the evolutionary mechanisms of their partners in the interaction. The final model in this section is a departure in that an individual distinguishes his close relatives from strangers and can choose a strategy when encountering strangers that differs from his strategy with close relatives. In this case, evolutionary stable behaviour maximizes the total fitness of a family. This result helps to explain the spread of cooperative and altruistic behaviour in 'prisoner's dilemma'-type interactions between relatives.

Let S denote the set of possible strategies; $\pi = (\pi_s, s \in S)$ is a distribution of individuals over strategies, while N_s and π_s are respectively the number and the share of individuals with strategy s, and $N = \sum_s N_s$ is the total size of the population. For any strategy s, fertility $fer_s(\pi, N)$ and viability $v_s(\pi, N)$ of the individuals playing this strategy depend on the distribution π and the size N in a given period. The basic model of evolutionary game theory is the replicator dynamics, RD. It assumes that every individual does not change his strategy during his life and new individuals inherit strategies from their parents. The evolutionary dynamics of such population satisfy:

$$N_s(t+1) = N_s(t) f_s(\pi(t), N(t)) \tag{1}$$

where $f_s(\pi, N) = fer_s(\pi, N) + v_s(\pi, N)$ is the fitness function of a strategy s.

Note that if the function takes the form $f_s(\pi, N) = a_s(\pi, N)\bar{f}(\pi)$, then system (1) implies the autonomous behaviour dynamics:

$$\pi_s(t+1) = \pi_s(t) f_s(\pi(t)) \Big/ \sum_\tau \pi_\tau(t) f_\tau(\pi(t)), \quad s \in S \tag{2}$$

that do not depend on $N(t)$.

The main content of the following results is that fitness turns out to be an endogenous payoff function for individuals in the given model. In particular, if a strategy provides greater fitness than another strategy under any distribution π, then the share of the worse strategy in the population tends to 0 as time tends to infinity. Moreover, if the strategy distribution $\pi(t)$ tends to some steady point π^* then only those strategies s that maximize fitness under π^* are included in this distribution with positive shares π_s^*.

In order to present the formal results, we need to recall or introduce several concepts of evolutionary game theory.

Population game, Nash equilibrium, dominance set of strategies

The model of a single interaction of individuals in a population at an instant or in a period of time is a population game (an analogue of a game in normal form in classical game theory). Formally, a population game G is given by a collection:

$$G = \langle S, f_s(\pi, \omega), s \in S, \pi \in \Pi, \omega \in \Omega \rangle \tag{3}$$

where S is the set of strategies of the participants of this game; $\pi = (\pi_s)_{s \in S}$ is the distribution of the players among strategies; $\Pi = \{\pi \mid \pi_s \geq 0, \sum_{s \in S} \pi_s = 1\}$ is a standard simplex; and $f_s(\pi, \omega)$ is the payoff function for players that use strategy s under strategy distribution π and other parameters of the model ω (for example total population size and environmental conditions). For biological populations, the payoff is usually taken to be fitness or a function that is consistent with fitness. For social populations, the payoff function usually corresponds to consumer utility, income, or profit. Consider now the main static optimality principles employed for the investigation of population interactions.

Definition 1 A *Nash equilibrium* for the population game of form (3) is a distribution π such that:

$$\forall \omega \in \Omega, \ \forall s \in S: \pi_s^* > 0 \Rightarrow s \in \operatorname{Arg} \max_{s \in S} f_s(\pi^*, \omega) \tag{4}$$

Assume that the payoff function for the game G has the following additive form:

$$f_s(\pi, \omega) = a(\pi, \omega) \bar{f}_s(\pi) + b(\pi, \omega), \quad a(\pi, \omega) > 0 \tag{5}$$

(In other words, the term that depends on the strategy chosen by a player is independent of the parameter ω.) Then, condition (4) is equivalent to condition (6), which does not involve ω:

$$\forall s \in S: \pi_s^* > 0 \Rightarrow s \in \operatorname{Arg} \max_{i \in s} \bar{f}_i(\pi^*) \tag{6}$$

If a game describes interaction in a biological population and the payoff function characterizes the fitness of the employed strategies, then the concept of Nash equilibrium completely corresponds to the Darwinian principle of natural selection: only the strategies that best fit a given distribution are present in π with positive probabilities (that is, survive).

The concept of Nash equilibrium was introduced in Nash (1951) and is the best-known optimality criterion used in strategic behaviour modelling. However, among Nash equilibria, there can exist unstable states that could not be realized in practice. For this reason, we also describe other optimality criteria.

Definition 2 Distribution π^* is a *strict equilibrium* for a population game G if:

$$\exists \varepsilon > 0, \exists s \in S \quad \text{such that } \pi_s^* = 1 \text{ and } \forall i \neq s, \omega \in \Omega$$
$$f_s(\pi^*, \omega) \geq f_i(\pi^*, \omega) + \varepsilon \tag{7}$$

For general payoff functions $f(\pi, \omega)$, Nash equilibria do not necessarily exist. Another optimality principle also associated with the Darwin's concept of natural selection is *dominance*.

Definition 3 Strategy s *dominates* strategy $i (s \succeq i)$ on the set of distributions $\Pi' \subseteq \Pi$ if:

$$\exists \varepsilon \geq 0: \forall \omega \in \Omega, \forall \pi \in \Pi', \quad f_s(\pi^*, \omega) \geq f_i(\pi^*, \omega) + \varepsilon \tag{8}$$

that is, for any distribution from the set Π', the strategy s provides a greater gain than does the strategy i.

Definition 4 The set $S' \subseteq S$ is called a *dominating set* if it can be obtained by iterative elimination of dominated strategies; that is:

$$S' = S_T \subset S_{T-1} \subset \cdots \subset S_1 = S$$
$$\text{where } \forall k \in \{1, \ldots, T-1\}, \ \forall i \in S_k \backslash S_{k+1} \ \exists s \in S_{k+1} \text{ such that } s \succeq i \text{ on } \Pi_k,$$
$$\Pi_k = \{\pi \in \Pi | \pi_s = 0, \forall s \notin S_k\} \tag{9}$$

The procedure described for iterative elimination of dominated strategies can be considered a quasi-dynamic model of behaviour microevolution within a population. Indeed, this procedure describes a sequential reduction of the set of strategies used by players: at each stage, more efficient (better-fitted) strategies are substituted for less-efficient ones.

If in the definitions we require the condition $\varepsilon > 0$, then s *strictly dominates* $i (s \succ i)$ and S is called a *strictly dominating set*.

The concepts of dominating by mixed strategies and a set dominating in mixed strategies are introduced in a similar way, with the payoff of the mixed strategy $f_\pi(\pi', \omega) = \sum_{s \in S} \pi_s f_s(\pi', \omega)$. (Domination is described in more detail in Fudenberg and Tirole, 1991.)

Stability of solutions

Special concepts of stability and asymptotic stability of a distribution over strategies consistent with the concept of Lyapunov stability are necessary for dynamic models of population behaviour. Consider these definitions in terms of Vasin (1995) for a dynamic model of the general form given by the system:

$$x(t+1) = \Phi(x(t)), \quad \pi(t) = h(x(t)) \tag{10}$$

Let $\pi(t, x^0)$ denote the distribution in the population over strategies at time t if the initial state of the system is x^0, and $\pi^0 = \pi(0, x^0)$ denote the corresponding distribution at the initial period. Distribution π^0 is called *stationary* for dynamic model (10) if:

$$\exists x^0 : \pi(t, x^0) \equiv \pi^0, \quad \forall t \geq 0 \tag{11}$$

that is, there exists an initial state of the system such that the distribution over strategies within the population remains invariant.

Distribution π^0 is called *stable* for dynamic model (10) if it is stationary and, in addition,

$$\forall \varepsilon > 0 \; \exists \delta > 0 : \forall x \notin O_\delta(x^0) \quad \pi(t, x) \in O_\varepsilon(\pi^0), \;\; \forall t \geq 0 \tag{12}$$

that is, for small deviations from the initial state, the distribution does not 'go far' from the initial one.

Distribution π^0 is called *asymptotically stable* for dynamic model (10) if it is stable and, in addition,

$$\exists \delta^0 > 0 : \lim_{t \to \infty} \pi(t, x) = \pi^0, \quad \forall x \in O_{\delta^0}(x^0) \tag{13}$$

that is, for sufficiently small deviations from the initial state, the distribution over strategies always tends to π^0.

For continuous-time models, the corresponding concepts are introduced similarly. Note that, for *autonomous models* in which $x = \pi$, these concepts of stability and asymptotic stability are equivalent to the ordinary definitions of Lyapunov stability and asymptotic stability (see Pontryagin, 1980).

Various theorems on the relation of static optimality principles with solution stability of RD are obtained in Taylor and Jonker (1978), Schuster *et al.* (1981), Vasin (1989) and Nachbar (1990). We give the formulations of these theorems following Vasin (1989, 1995).

Theorem 1 (on the relation of Nash distributions with stable points of RD). Assume that the fitness function f_s is representable in the additive

form (5). Then, the following propositions are valid:

(1) any stable distribution π of system (1) is a Nash equilibrium in the population game $G = \langle S, f_s(\pi, N), s \in S \rangle$
(2) if for a certain path $\bar{N}(t)$, the initial distribution $\bar{N}(0) > 0$ and $\exists \lim_{t\to\infty} \pi(\bar{N}(0), t) = \pi^*$, then π^* is a Nash equilibrium of the specified population game.

Theorem 2 (on asymptotic stability of a strict equilibrium). Assume that π is a strict equilibrium of the population game $G = \langle S, f_s(\pi, N), s \in S \rangle$. Then π is an asymptotically stable distribution of system (1).

Theorem 3 (on the relation between dominating sets of strategies and behaviour dynamics). Assume that $\bar{S}$ is a strictly dominating set of strategies in the game $G' = \langle S, \ln f_s(\pi, N), s \in S \rangle$. Then, for any $s \notin \bar{S}$ and any $\bar{N}(0) > 0$, $\lim_{t\to\infty} \pi_s(\bar{N}(0), t) = 0$ on the corresponding path of system (1).

Note that in the case of dominance by pure strategies, theorem 3 is also valid for the standard population game $G = \langle S, f_s(\pi, N), s \in S \rangle$. However, in the case of dominance by mixed strategies (that is, by distributions), it is necessary to consider the logarithm of fitness as a payoff function, because there exist examples where a strategy strictly dominated by a distribution in the game $G = \langle S, f_s(\pi, N), s \in S \rangle$ is not eliminated on the trajectories of the discrete-time RD.

Many studies consider a continuous version of RD (see Taylor and Jonker, 1978; Hofbauer and Sigmund, 1988, and others). The population size in this model varies continuously. Function $f_s(\pi, N)$ determines the difference between the rates of fertility and mortality in a group of individuals employing strategy $s \in S$, whereas other assumptions are similar to the discrete variant. As a result, the dynamic equations take the form $\dot{N}_s = N_s f_s(\pi, N)$ and, for functions in the additive form (5), by changing the time $d\tau = a(\pi, N)dt$, one obtains the autonomous system:

$$\dot{\pi}_s = \pi_s \left(\bar{f}_s(\pi) - \sum_{i\in S} \pi_i \bar{f}_i(\pi) \right)$$

Analogues of the theorems presented above are also proved for these models (see, Bomze, 1986; Weibull, 1996; Taylor and Jonker, 1978; Samuelson and Zhang, 1992).

Of course, these results strongly depend on the evolutionary mechanism of replicators. As an alternative, consider a model of random imitation where a newborn follows a strategy of a randomly chosen adult. Then the population

dynamics are described by:

$$N_s(t+1) = N_s(t)v_s(t) + \sum_r N_r(t)fer_r(t)\frac{N_s(t)v_s(t)}{\sum_r N_r(t)v_r(t)}, \quad s \in S$$

Note that the right-hand side is of the form $N_s(t)v_s(t)V(t)$ where $V(t)$ does not depend on strategy s. Under general assumptions, such dynamics are coordinated with the function $v_s(t)$ in the sense of theorems 1–3. Thus, viability (instead of fitness) turns out to be an endogenous payoff function of individuals in the corresponding dynamical process.

A model of evolutionary mechanism selection

Proceeding from the previous example it seems that we have exchanged arbitrariness in the choice of payoff functions for arbitrariness in the choice of evolutionary mechanisms. However, actual evolutionary mechanisms are subject to natural selection. Only the most efficient mechanisms survive in the process of competition.

Consider the corresponding model of a society including several populations that differ only in their evolutionary mechanisms. Individuals of all populations interact and do not distinguish population characters in this process. Thus, the evolutionary mechanism of an individual is an unobservable characteristic. Fertility and viability functions $fer_s(\pi, N)$, $v_s(\pi, N)$ describe the outcome of the interaction and depend on the total distribution over strategies and the size of the society. The set of strategies S and the functions are the same for all populations. Let L denote the set of populations, N_l denote the size of population l, $\pi^l = \{\pi^l_s, s \in S\}$ be a distribution over strategies in population l. Then the total distribution over strategies is $\pi = \sum_l (N^l/N)\pi^l$. Assume that operator Φ^l corresponds to the evolutionary mechanism of population l and determines the dynamics of distribution π^l. Then the dynamics of the society are governed by:

$$N^l(t+1) = N^l(t)\sum_s \pi^l_s(t) f_s(\pi(t), N(t)) \tag{14}$$

and

$$\pi^l(t+1) = \Phi^l(\pi^k(t), N^k(t), k \in L) \tag{15}$$

where $f_s(\pi, N)$ is a fitness function of strategy s.

Theorem 4 Let there exist a population of replicators in the society. Then the total distribution over strategies meets the following analogues of theorems 1 and 2:

(1) any stable distribution π of system (14) is a Nash equilibrium of the population game $G = \langle s, f_s(\pi, N), s \in S \rangle$;
(2) if initial distribution $\pi_0 > 0$ and $\exists \pi = \lim_{t\to\infty} \pi(\breve{N}_0, t)$ then π is a Nash equilibrium of the population game G;
(3) if π is a strict equilibrium of the game G then π is an asymptotically stable distribution for system (14, 15).

Proof of Theorem 4 Let us show that convergence of $\pi(t)$ to π implies that π is a Nash equilibrium. Suppose on the contrary that $\pi_s > 0$ for some non-optimal strategy $s \notin \text{Arg max}_{i\in S} f_i(\pi)$. Consider a neighbourhood of π. Due to the continuity of f_i, the gain from strategy s in a small enough vicinity $O_\delta(\bar{\pi})$ is still strictly less than for any optimal strategy $r \in \text{Arg max}_i(f_i(\pi))$. Moreover, for any mixed strategy $d \in O_\delta(\pi)$, $\bar{f}_r(\pi') > \bar{f}_d(\pi')$ under $\pi' \in O_\delta(\pi)$.

In a population of replicators consider those individuals who apply strategy r, which is the optimal response to π. The growth rate of this group is higher than the average growth rate for the overall society and the share of this group in society will increase until distribution π leaves vicinity $O_\delta(\pi)$. We get a contradiction with the assumption that π is a limit point for the positive initial distribution over populations and strategies. The remaining propositions are proved similarly.

The generalization of theorem 3 for the elimination of dominated strategies is possible under stricter assumptions on the variety of evolutionary mechanisms. For any evolutionary mechanism Φ and a pair of strategies s, r, let us call an s, r-substitute of mechanism Φ a mechanism $\Phi_{s,r}$ such that for strategies other than s and r the shares of individuals who apply these strategies change as under mechanism Φ, except that instead of strategy s they always play r. According to (Vasin, 1995), if for any s, r, l the set of mechanisms includes all possible substitutes $\Phi^l_{s,r}$, then $\pi_s(t) \to 0$ as $t \to \infty$ for any strictly dominated strategy s. Thus, the evolutionary mechanism selection model confirms that individual fitness is an endogenous utility function for self-reproducing populations.

Does this result correspond to actual behaviour? First, consider behaviour in biological populations. A general opinion of biologists is that, on the whole, the principle of individual fitness maximization does not contradict actual behaviour (ESEB, 1991). However, there are well-known examples of altruistic and cooperative behaviour that apparently do not correspond to this principle. The concepts of altruistic and cooperative behaviour may be illustrated by different variants of the prisoner's dilemma (see Owen, 1974). In this symmetric two-player game each player has two possible strategies: to cooperate (C) or to be selfish (S). As an example we consider the following

payoff matrix:

	C	S
C	$(5, 5)$	$(1, 6)$
S	$(6, 1)$	$(2, 2)$

In the general prisoner's dilemma, given any behaviour of the other player, the selfish strategy is more profitable $u_{SS} > u_{CS}, u_{SC} > u_{CC}$, while at the same time the total gain is maximal when both players cooperate $u_{CC} > (u_{CS} + u_{SC})/u_{SS} > u_{CC}$. In this game there exists a unique Nash equilibrium, which corresponds to selfish behaviour and is also a dominance solution. However, in actual prisoner's dilemma-like situations, players often cooperate.

Altruistic behaviour deviates even more from individual fitness maximization. Consider the following payoff matrix (A denotes altruistic behaviour, S denotes selfish behaviour):

	A	S
A	$(5, 5)$	$(1, 10)$
S	$(10, 1)$	$(2, 2)$

Here, altruistic behaviour by one player combined with selfish behaviour by the other player corresponds to total fitness maximization. Meanwhile, the altruist in such an outcome obtains less than the guaranteed payoff he could get under his Nash equilibrium strategy.

As examples of cooperative behaviour in biological populations, consider the behaviour of animals that take turns standing guard, or predators that participate in joint hunting. As for altruism, it may be observed in interaction between relatives. The altruistic behaviour of parents towards children is rather widespread and does not contradict the concept of fitness maximization, since fitness is equal to the sum of fertility and viability. More interesting examples of altruism are those that do not relate to individual fitness maximization; for instance, the behaviour of social insects. The explanation is that individuals in the families of social insects are close relatives. A shortcoming of the model of direct inheritance is that it only takes into account the 'parent–child' relationship but does not consider relations between siblings, cousins, and so on. Taking these relations into account, it is possible to explain the spread of cooperative and altruistic behaviour in the sense of total fitness maximization for the group of relatives

Let us describe the corresponding model. As above, assume that interaction in a population is characterized by the set of strategies S. But in contrast to the previous models, individuals can distinguish between siblings ('sibs'), that is brothers and sisters, and other members of the population, and choose a strategy based on this characteristic. Thus, a full strategy $(s, s') \in S \times S$

includes a component s which is applied to sibs and s' for other individuals (strangers). At a time period, an individual interacts with sibs with some frequency $\lambda_r \in (0, 1)$ and with strangers with frequency $1 - \lambda_r$. Individual fitness additively depends on the results of interactions with relatives and the rest of the population:

$$\bar{f}_{s,s'}(\pi') = \lambda_r f_s^r(s) + (1 - \lambda_r) f_{s'}(\pi')$$

where f_s^r and $f_{s'}$ determine the results of the interaction respectively with relatives and strangers, while π' is a distribution over component s'. Following the RD model, we assume that all sibs play the same strategies. Thus, the interaction is characterized by a population game:

$$\overline{G} = \langle \overline{S} = \{(s, s') \in S \times S\}, \bar{f}_{s,s'}(\bar{\pi}) = \lambda_r f_s^r(s) + (1 - \lambda_r) f_{s'}(\pi') \rangle$$

where $\bar{\pi}$ is a distribution over full strategies.

Theorem 5 Any strategy (s, s') where $s \notin \text{Arg max}_i f_i^r(i)$ is strictly dominated by strategy (s^*, s') where $s^* \in \text{Arg max}_i f_i^r(i)$. Distribution $\bar{\pi}$ is a Nash equilibrium if for all specified non-optimal strategies $\pi_{ss'} = 0$ and the corresponding distribution π' is a Nash equilibrium of the game $\langle S, f_{s'}(\pi') \rangle$.

Thus, in any conflict similar to a prisoner's dilemma, sibs play the cooperative strategy with respect to each other. In order to explain altruistic behaviour, consider the following modification of the model. Assume that in the interaction sibs may take on one of two different roles (for instance, an older individual α and a younger one β). Let the strategy and the fitness function depend on the role. Then the full strategy with respect to relatives is determined by the pair $s = (s_\alpha, s_\beta) \in S \times S$. Only strategies $s^* = (s_\alpha^*, s_\beta^*) \rightarrow \max_{(s_\alpha, s_\beta)} (f_{s_\alpha}^\alpha(s_\beta) + f_{s_\beta}^\beta(s_\alpha))$, which provide the maximum total fitness, survive elimination of strictly dominated strategies. Proceeding from theorems 4 and 5, we may conclude that evolution in self-reproducing populations leads to behaviour that maximizes the total fitness of sibs. Although the last model contains an implicit restriction on the evolutionary mechanism that determines the distribution over strategies with respect to relatives, this restriction is not important: any other mechanism would lose in competition with the given mechanism, which realizes the optimal strategy s^* specified in theorem 5. (For the earlier example, either the older sib acts as an altruist with respect to the younger one, or vice versa.)

Note that these results may be generalized for relations between cousins, second cousins, and so on. Then a full strategy includes variants of behaviour $s_1, \ldots, s_k$ with respect to relatives of different degrees $1, \ldots, k$ and s' with

respect to strangers. The fitness function is of the form:

$$f_s(\pi) = \sum_i \lambda_i f_{s_i}(s_i) + (1 - \sum_i \lambda_i) f_{s'}(\pi')$$

where $\lambda_i \in (0, 1)$ characterizes the frequency of interaction with relatives of degree i. As above, we assume that all relatives apply the same strategy with respect to each other. This assumption corresponds to the model of direct inheritance and does not restrict generality within the context of the proposed model of evolutionary mechanism selection. Then, elimination of strictly dominated strategies leads to optimization of relations between relatives: surviving strategies are those s^* such that $s_i^* \in \text{Arg max}_{s_i} f_{s_i}(s_i)$, $i = 1, \ldots, k$.

In many biological populations, as well as in human populations with low migration, interacting individuals usually have common ancestors seven generations earlier. Consider a simple model that justifies this proposition. Let two individuals be randomly chosen from a population of the size 10^6. Assume that they have no common ancestors in the previous seven generations. Then seven generations ago each individual has 64 ancestors of every sex. Let the total size of the population at that time be 10^4 that corresponds to the average number of children per family being about 4. A necessary condition for the individuals to be non-relatives is that no female-ancestor of one individual married a male-ancestor of the other individual. The probability of such an event is $(1 - 64/5000)^{128} < 0,\ 1$.

According to the above results, the ancestral linkages should lead us to expect fairly widespread cooperative and altruistic behaviour that maximizes total fitness. However, actual behaviour often does not correspond to these expectations. Examples of severe competition between close relatives (some species even eat up the offspring of relatives) are well-known. One of the reasons for the limited prevalence of cooperative behaviour is that widespread cooperation is not invasion-proof against selfish mutants.

Imagine a population with cooperative behaviour where all individuals play strategy s^* that maximizes total fitness. Consider the following mutation: there appears an individual that maximizes his individual fitness by playing $s' \to \max_s f_s(s^*)$. In the first generation such an individual will garner a reproductive advantage because all other members will cooperate with him while he behaves in a selfish manner. In the second generation the mutant's offspring would be in a somewhat less-favourable position, as they will occasionally interact with each other. In the general case, the outcome depends on the relationship among the coefficients λ_i. The share of the selfish mutants in the population will increase until their interactions turn out to be with other selfish mutants, to a sufficiently large extent. Thus, the total fitness-maximizing strategy is not invasion-proof. Another difficulty is that an individual may be unable to distinguish between his cousins and more distant relatives and strangers.

Note that the prevalence of cooperative and altruistic behaviour has been studied in many papers (see Axelrod, 1984, and the references given therein). For instance, the theory of repeated games explains the prevalence of cooperation in repeated conflict situations proceeding from individual fitness optimization (see Van Damme, 1987; Vasin, 1997). Repetition provides the possibility of punishing those individuals who fail to behave cooperatively. Taking into account such future punishments, cooperative behaviour turns out to be individually profitable. The evolutionary models considered above do not assume that an individual takes part in repeated interaction and obtains information about the previous behaviour of his partners. Cooperative behaviour among relatives can nevertheless spread broadly due to the selection of evolutionary mechanisms.

3 Peculiarities of behaviour evolution in human populations

Actual behaviour in modern populations maximizes neither individual nor group fitness. Russia, where reproductive conditions have been unfavourable during the last decade, is a special case. Let us take social-welfare states such as Sweden or Germany. Favourable conditions for survival and normal biological development are guaranteed there to any newborn citizen regardless of his or her social origin. However, demographic statistics indicate that most people ignore the advantages in reproductive opportunities.

For instance, the pool of Stockholm inhabitants in 1995 showed that 70 per cent of the adult population had no children and did not plan to have them. The fertility rate in Sweden and Germany is almost as low as in Russia: respectively 10 and 9 per 1,000. According to the demographic forecast by the UN, Germany's population will decrease from 82 million in 1999 to 80 million in 2025 (United Nations, 1996: 352–3).

This situation is quite different from the story told by Maynord Smith (1984) about two kinds of reproductive strategies. According to his classification, individuals of some species maximize the quantity of the offspring while the others maximize their quality measured by the probability of surviving and hold the reproductive capacity up to the reproductive age. In our case the behavioural strategy does not provide for simple reproduction since the average number of children per individual is less than one. Moreover, the individual spends a large part of his or her resources on purposes that are unrelated to his or her survival and reproduction.

Why do the evolutionary models examined above seem not to apply directly to current social populations? Three factors can explain this discrepancy:

(1) The most important cause of the disjunction between the models and reality is the impossibility of children reliably inheriting parental

strategies within social populations. Behaviour strategies are so complicated and the environment so variable that an individual could spend most of his life trying to teach his descendants and would typically not succeed. The division of labour and differential access to educational institutions appeared at an early stage of mankind's development, and such institutions play a crucial role in moulding the behaviour of new generations (Moiseev, 1999). Later, governments, churches and other organizations realized the importance of educational institutions and effectively used them to mould desirable behaviour. In any country children and young people are taught to obey the state's laws, irrespective of their impact on individual reproduction. Encouragement to report on the political views of family members under some totalitarian regimes, and education in accordance with a 'one family–one child' policy, are obvious examples from the history of authoritarian governments. Private capital uses more subtle means for regulating behaviour (see factors 2 and 3 below) but also influences individual reproduction. For instance, data on Russia show that the fertility rate for employees of large corporations is considerably lower than the average rate for the whole population.

(2) The problem of payoff evaluation for different behaviour strategies is rather complicated even for theoretical formulation, and in practice decision-makers are usually unable to carry out a complete evaluation within a reasonable period of time. Biological evolution, however, has generated various mechanisms that facilitate fast decisions that are generally close to optimal in terms of fitness.

One such mechanism is the feeling of pleasure or satisfaction. In nature, actions that are pleasant or conducive to giving pleasure are usually rational in terms of individual reproduction. In particular, the provision and consumption of food and some other resources are necessary for reproduction. However, devotion to this objective function may impede reproduction in some instances. One example of fitness-impeding commitment to food consumption within an ecological system is the interaction between a lamehuza beetle and ants who feed on secretions produced by the beetle. Sometimes the secretions have a drug-like effect on the ants: they throw the queen out of the ant-hill, put the lamehuza beetle in the queen's place and take care of it in order to get more secretions. In time this ant family perishes.

This example is exceptional for ecosystems, where excessive or harmful consumption is atypical. The situation in modern human society is different. A lot of people consume large amounts of alcohol, tobacco, excessive food and other goods that are harmful or at least unnecessary for reproduction. There are many families and individuals spending their

entire lives to earn money for such consumption and finally having at most one child.

(3) Another auxiliary mechanism of strategy determination is imitation. It promotes learning and permits coordination of the actions of individuals within groups and thus increases their fitness. At the same time imitation creates an additional possibility for behaviour manipulation by choosing an appropriate leader.

The methods of behaviour manipulation described above (control over the educational process, use of pleasure incentives and the imitation mechanism) have been practiced since ancient times. But the situation has dramatically changed with the development of mass media in the twentieth century. Currently, television had the potential of influencing billions of people by playing the role of a teacher and creating models for imitation.

Now consider who or what changes the objective functions in order to promote certain types of behaviour in social populations. In our previous examples, we identified two variants. In the case of the ants and the lamehuza beetle, the manipulator is an individual belonging to another population. The interaction between the populations is of the 'predator–prey' type (see Volterra, 1931). In the case of social insects, individual behaviour is formed by a self-reproducing *superindividual* – the family of ants, bees or other social insects. While suppressing individual reproduction of some part of the population, the evolutionary mechanism provides efficient reproduction of these superindividuals and of the population as a whole.

In general a superindividual related to a given biological or human population is a self-reproducing structure that includes some individuals of this population among elements. Apart from individual organisms, the superindividual may comprise other material or immaterial components. In social populations we meet similar and more complicated variants of behaviour regulation. In addition to superindividuals of a *biological nature* (families), there exist self-reproducing superindividuals of a *social-economic nature*; for instance, corporations, art and scientific schools, public and religious organizations, and government institutions. Instead of replication, superindividuals may either grow, involving new human and material resources, or collapse.

Thus, superindividuals use the population as a resource for their own reproduction and growth, and influence population behaviour for this purpose. The more intensive is the specific activity of involved individuals, the faster is growth and development of the superindividual. Since the time and energy of each person is limited, the superindividual often (usually) suppresses other kinds of activity, in particular, individual reproduction. The relevant theory of social and economic behaviour should reflect the processes of interaction and selection of self-reproducing individuals at different levels. Proceeding

from the results of section 2, I conjecture that individual utility functions in modern society are determined by this process: they induce behaviour that provides reproduction and the growth of certain superindividuals. Some individuals understand this relation. As a result, a person consciously distributes his time and energy among activities that concern reproduction of his family, local and professional communities, nation and other superindividuals within his sphere of interest. In other cases, an individual is not aware of his role in the evolutionary process and acts according to some auxiliary utility function or mechanism formed by this process. But, as in the previous case, his activity serves the reproduction of some superindividual(s).

National populations and superindividuals have coexisted for hundreds and thousands years. The suppression of individual reproduction might even be useful for the growth and development of the whole. However, one unfortunate tendency in several European countries now is that the population is declining at a substantial rate. According to the UN demographic forecast, the population in Eastern Europe and Russia will decrease by one-third by 2050. The Spanish population will decline from 40 to 30 million (see *El Pais*, 2002, on the UN 2050 population forecast). The role of private and government institutions in this process requires careful study.

4 Conclusion

This chapter has discussed models of individual utility function formation in the evolution of population behaviour, demonstrating that if the replication of behaviour strategies is possible then an endogenous utility function is the rate of reproduction, that is the sum of the fertility and viability of individuals with a given strategy. For interactions among close relatives, evolution selects strategies that maximize the total rate of family reproduction.

However, such behaviour is atypical for modern human populations, where complete replication is actually impossible. At the same time, human behaviour often fails to correspond to the standard economic concept of *homo economicus*. My conjecture is that this behaviour is determined by an evolutionary process. However, competition and the selection of strategies takes place not only at the individual level, but also at the level of superindividuals. They are self-reproducing objects that include people as elements of their structure and can influence behaviour dynamics. Besides the government, political and religious organizations and large corporations are highly influential superindividuals in the modern world. Mass media and educational institutions also impact on population behaviour.

References

Axelrod, R. (1984) *The Evolution of Cooperation* (New York: Basic Books).

Bomze, I.M. (1986) 'Non-cooperative Two Person Games in Biology: A Classification', *International Journal of Game Theory,* vol. 15, pp. 31–57.

El País (2002) 'UN 2050 Population Forecast 31.3 million while State Bureau Puts Figure at 41.2', 3 December, p. 3 (in Spanish).
ESEB (1991) *Third Congress of the European Society for Evolutionary Biology, Abstracts* (Debrecen: European Society for Evolutionary Biology).
Friebel, G. and S. Guriev (2000) 'Why Russian Workers Do Not Move: Attachment of Workers through In-Kind Payments', CEPR Discussion Paper no. 2368.
Fudenberg, D. and J. Tirole (1991) *Game Theory* (Cambridge MA: MIT Press).
Germeyer, U.B. (1976) *Igri s neprotivopoloynimi interesami* (*Games with Non-Antagonistic Interests*) (Moscow: Nauka) (in Russian).
Hofbauer, J. and K. Sigmund (1988) *Dynamical Systems and the Theory of Evolution* (Cambridge: Cambridge University Press).
Maynard Smith, J. (1982) *Evolution and the Theory of Games* (Cambridge: Cambridge University Press).
Moiseev, I. (1999) *Byt' ili ne byt' chelovechestvu?* (*To Be or Not to Be for Mankind*) (Moscow: Nauka) (in Russian).
Myles, G. (1995) *Public Economics* (Cambridge: Cambridge University Press).
Nachbar, J.H. (1990) 'Evolutionary Selection Dynamics in Games: Convergence and Limit Properties', *International Journal of Game Theory*, vol. 19, pp. 59–89.
Nash, J. (1951) 'Non-Cooperative Games', *Annals of Mathematics*, vol. 54, pp. 286–95.
Ok, Efe A. and F. Vega-Redondo (2001) 'On the Evolution of Individualistic Preferences: An Incomplete Information Scenario', *Journal of Economic Theory*, vol. 97(2), pp. 231–54.
Owen, G. (1974) *Game Theory* (Philadelphia: W.B. Saunders).
Pontryagin, L.S. (1980) *Differentsial'nye uravneniya* (*Differential Equations*) (Moscow: Nauka) (in Russian).
Samuelson, L. (2001) 'Introduction to the Evolution of Preferences', *Journal of Economic Theory*, vol. 97(2), pp. 225–30.
Samuelson, L. and J. Zhang (1992) 'Evolutionary Stability in Asymmetric Games', *Journal of Economic Theory*, vol. 57, pp. 363–91.
Schuster, P. and K. Sigmund (1983) 'Replicator Dynamics', *Journal of Theoretical Biology*, vol. 100, pp. 1–25.
Schuster, P., K. Sigmund, J. Hofbauer and R. Wolf (1981) 'Self-regulation of Behavior in Animal Societies. Games between Two Populations without Self-interaction', *Biology and Cybernetics*, vol. 40, pp. 9–15.
Slinko, I.A. (1999) 'Multiple Jobs, Wage Arrears, Tax Evasion and Labor Supply in Russia', Working Paper no. BSP/99/018 (Moscow: New Economic School).
Taylor, P. and L. Jonker (1978) 'Evolutionary Stable Strategies and Game Dynamics', *Mathematical Biosciences*, vol. 40, pp. 145–56.
United Nations (1996) *World Population Prospects: The 1996 Revision*. Annex II & III: *Demographic Indicators by Major Area Region and Country* (New York: United Nations).
Van Damme, E. (1987) *Stability and Perfection of Nash Equilibrium* (Berlin: Springer).
Vasin, A. (1989) *Modeli kollektivnogo dinamiki povedeniya* (*Models of Collective Behavior Dynamics*) (Moscow: Moscow University Press) (in Russian).
Vasin, A. (1995) 'On Some Problems of the Theory of Collective Behaviour', *Obozrenie prikladnoy i promishlennoy matematiki,* vol. 2, pp. 1–20 (in Russian).
Vasin, A. (1998) 'The Folk Theorems in the Framework of Evolution and Cooperation', Interim Report no. IR-98-074, International Institute for Applied Systems Analysis, pp. 1–8, Laxenburg, Austria.
Volterra, V. (1931) *Leçons sur la théorie mathématique de la lutte pour la vie* (Paris: Gauthier-Villars).
Weibull, J. (1996) *Evolutionary Game Theory* (Cambridge MA: MIT Press).

6

Moral Hazard, Contracts and Social Preferences: A Survey*

Florian Englmaier
University of Munich, Germany

1 Introduction

This chapter provides a non-technical survey of recent contributions to the emerging field of behavioural contract theory that try to incorporate social preferences into the analysis of optimal contracts in situations of moral hazard. The presence of these social preferences is confirmed by numerous studies. Taking them into account when analysing optimal contracts generates important new insights, and might help us gain a better understanding of real-world contracts and organizational structures.

A central question that economists have been facing for a long time is how to give workers the right incentives to motivate them to perform as desired by the Principal. Over the years, the moral hazard problem has become one of the most intensely analysed. As a result, many insights have been gained and the problem also seems to be one of the best understood in economics. Having said this, the theory has an important shortcoming. Real-world contracts seldom look like those predicted by theory. Often contracts are linear and simpler, incentives are sometimes more high-powered or the wage schedule more compressed than expected. And some features, such as the widespread use of employee stock-option plans, seem somewhat bewildering.

One reason for this shortcoming may be that economic theorists have based their models on the assumption that the agent is a solely self-interested *homo economicus*. Although this is often a good working assumption, in the specific context of labour relations it misses out on some important aspects like social ties, team spirit or work morale, which appear fundamental to researchers and practitioners in the field of human resources. With some notable exceptions, this gap between economic theory and research in human resources is only now beginning to close.

* I am indebted to my colleagues Ingrid Königbauer, Markus Reisinger and Astrid Selder for their valuable comments and suggestions and also to the editors of this volume for their inputs. An earlier version of this chapter was presented at the 13th World Congress of the International Economic Association in Lisbon in September 2002.

Kandel and Lazear (1992), in an early theoretical paper, try to incorporate social relations into a formal model. They model 'peer pressure' where co-workers inflict social sanctions on agents who fall short of some norm. As an additional instrument to provide incentives, peer pressure is efficiency-enhancing, and this can have implications for a firm's policy. Kandel and Lazear highlight the importance of profit-sharing plans as well as 'spirit-building activities' as means of enhancing the power of peer pressure. Similarly, Rotemberg (1994) examines whether it would be optimal to develop altruistic preferences in work relationships. In his model, agents can choose whether to be altruistic towards their co-workers. Although intuitively that never seems to be an optimal thing to do, it may in fact be beneficial, since altruism gives commitment power. In a team production setting with strategic complementarities, they can now commit to exerting a high level of effort as it is now in their best interest to do so. Hence in such settings the efficient outcome can be realized if agents can choose to become altruistic beforehand. It may thus be good for firms to give their workers the chance to develop altruistic feelings towards each other, such as by socializing a lot.

All these papers use a somewhat *ad hoc* specification of not solely self-centred preferences. But recently experimental and field evidence has helped to amend the standard utility function and move it to a sounder footing, and to develop extensive form models of social preferences. Further below we will discuss several of these amendments. However, Fehr and Schmidt's (1999) model serves as a reference point in most of the papers presented in this survey.

The present chapter is modest in scope and only addresses the theoretical contributions to the moral hazard problem. It does not address experimental work on incentive provisions,[1] nor address other informational problems such as adverse selection. It is structured as follows: section 2 spells out why social preferences can add valuable insights to the analysis of incentive provisions and how to model these social preferences; section 3 analyses the standard one-agent-one-principal problem, as studied by Holmström (1979), our exposition following Englmaier and Wambach (2002); section 4 then turns to a special case of multi-agent settings – tournaments; while section 5 deals with team production problems. The concluding section 6 outlines some promising topics for future research.

2 Social preferences: evidence and modelling approaches

Akerlof (1982) was probably the first to point to the importance of social preferences for labour market outcomes in a theoretical model. He characterized labour relations as a form of gift exchange. In a situation where we cannot contract effort, the employer offers the employee a generous wage, hoping that the employee will reciprocate this 'gift' with more than minimum effort.

In a subsequent paper, Akerlof and Yellen (1988) argue that the resulting market clearing wage may account for equilibrium unemployment.

These arguments are experimentally backed by two papers: Fehr, Kirchsteiger and Riedl (1993) and Fehr, Gächter and Kirchsteiger (1997). In experiments, these authors replicate labour markets and confirm the results of Akerlof (1982) and Akerlof and Yellen (1988). In their data they show that there is a positive relation between wage offers by firms and work effort responses by workers. And firms seem to understand the possibility of triggering effort by this means, since they make deliberate and extensive use of it. As a result, even in competitive double-auction environments, the wage level remains above the market-clearing level, resulting in involuntary unemployment.[2]

There is a great deal more experimental evidence on the importance of social preferences for incentive provisions (see references in Fehr and Falk, 2002; Fehr and Schmidt, 2003; and Gächter and Fehr, 2002). Important additional evidence is also provided by Bewley (1999) who undertook an extensive survey. He asked a large number of managers their opinions on wage cuts and other pay policies. These interviews clearly highlight that managers fear a breakdown of working morale if they make use of an adverse labour market situation in an 'unfair' manner and cut wages.

Given this evidence, it is no surprise that there have been several attempts to amend standard theory with social preferences. Rabin (1993) tries to incorporate fairness into game theory, and in his model of a static simultaneous move game, individual utility depends on a belief in the other's intentions. If you believe that your opponent wants to do something in your favour, your utility increases by returning this favour. If, however, you believe that your opponent will hurt you, the optimal response is to retaliate. One can easily see that there are generally multiple equilibria sustained by self-fulfilling prophecies. A good one, where each believes that the opponent has good intentions and where these expectations are met in equilibrium, and a bad one where each player believes the other to be evil-minded and this belief again is met in equilibrium. Dufwenberg and Kirchsteiger (2004) extend Rabin's paper to sequential games and Falk and Fischbacher (1998) is another attempt at a general model.

The equilibrium predictions of these models crucially depend on a player's belief about the other player's intentions. This is tricky to deal with and inherently hard to test empirically or experimentally. Therefore models have been developed that try to capture social preferences while only placing observable variables, such as monetary outcomes, as conditions.

Generally, in these models there is a separable term added to standard utility which captures relative income comparisons. Agents suffer a utility loss if they do not get their 'fair' share of total output, that is, if the allocation is 'inequitable'. For most experimental settings inequity can be replaced

by inequality. The two most prominent models are those by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999).

While in the Bolton and Ockenfel (2000) model agents compare their own payoff to the average payoff of all other agents, in Fehr and Schmidt (1999) the relative income comparison is a weighted sum of comparisons between each agent separately. Note that with only two agents those two models coincide. Although both models are close to one another in spirit, somehow the Fehr and Schmidt (1999) model has proved to be a bit more flexible and most of the models discussed in this survey use it, or a variant of it, as a reference point. A more detailed exposition of this model thus seems appropriate.

Fehr and Schmidt assume a utility function of the following form:

$$U_i(x_i) = x_i - \alpha_i \frac{1}{(n-1)} \sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{(n-1)} \sum_{j \neq i} \max\{x_i - x_j, 0\}$$

That is, utility is additively separable in income and disutility from inequitable outcomes.

The first addend x_i is standard and depicts utility from monetary payoff. The second addend, $\alpha_i(1/(n-1)) \sum_{j \neq i} \max\{x_j - x_i, 0\}$, creates disutility whenever the agent's payoff falls short of another player's payoff, whilst the third addend, $\beta_i(1/(n-1)) \sum_{j \neq i} \max\{x_i - x_j, 0\}$, reduces utility when the reverse holds true, that is when the agent is better off than an opponent.

The parameters α_i and β_i denote the weight that the agent puts on those social comparisons. The following restrictions are placed on these parameters: $\alpha_i \geq \beta_i$ and $\beta_i \in [0; 1]$. This implies that agents suffer more from being worse-off than others than from being better-off. And the assumption that $\beta_i \in [0; 1]$ rules out both 'status-seeking' and situations where agents would forego own material payoff in order to reduce favourable inequity.

This functional form depicts 'self-centred inequity-aversion', that is, agents are not really interested in the allocation of wealth in the population, they are only interested in their relative standing in this wealth distribution. Although the aversion towards disadvantageous inequity is more pronounced than aversion towards advantageous distributions, Fehr and Schmidt need both parts of the inequity aversion to explain observed behaviour. Moreover, they allow for heterogeneity in the population and inequity aversion still has relevance even if a substantial part of the population is purely self-interested.

While intention-based models clearly provide a more realistic depiction of reality, they are highly complicated to deal with.[3] Even very simple and abstract experimental games are hard to solve and the more interesting problems basically become intractable. Thus the purely outcome-based models serve as shortcuts for modelling reciprocal preferences. While they

are still analytically tractable they capture many aspects of reality and do a remarkably good job in explaining experimental evidence.

3 The moral hazard problem

As already highlighted in the introduction, the moral hazard problem is one of the central problems of labour market analysis. Englmaier and Wambach (2004)[4] were the first to introduce inequity aversion into moral hazard theory, and they amended Holmström's (1979) seminal paper. In this model one principal and one agent interact. The agent's (unobservable) choice of effort influences the distribution of profits. Englmaier and Wambach make one important change in Holmström's model: the agent's preferences exhibit inequity aversion as s/he compares herself/himself to the Principal. As this model was the first of its kind, more emphasis will be placed on its exposition in what follows.

The agent's utility is given by:

$$U_A = u[w(x)] - c(e) - \alpha G\{[x - w(x)] - w(x)\}$$

Utility consists of three parts. Of these, two parts are standard, namely, $u[w(x)]$, the utility derived from monetary income, and $c(e)$, the disutility from effort. For these two parts the standard assumptions apply, that is utility increases with income (although agents may be risk-averse or risk-neutral) and the costs of effort increase. The third part is new, $\alpha G\{[x - w(x)] - w(x)\}$. This captures the disutility from inequitable outcomes where α is the weight the agent puts on achieving equitable outcomes.

The convex cost function $G(\cdot)$ displays the disutility from inequity. It is assumed to be zero at $x - w(x) = w(x)$, that is for equitable outcomes where the agent's wage payment $w(x)$ equals the Principal's net profit $[x - w(x)]$, and also to be flat at this point. But the further away the equity outcome is, the more marginal disutility increases. A quadratic function would do that job. However, $G(\cdot)$ is not required to be symmetric around zero, that is agents may, quite realistically, suffer a lot more from being worse off than from being better off than the Principal. This convexity implies an aversion against lotteries over different levels of inequity.

Note that the authors assume that agents will only compare gross payments. Their results would hold qualitatively if agents compare rents, that is, payoffs net of effort costs, but the exposition of the results would be more cumbersome. Furthermore, it is assumed that the Principal is of a standard type, that is, s/he is just interested in her/his expected payoff and not in relative comparisons. Again, all the results would go through qualitatively but the exposition would be more cumbersome.

Thus the Principal's problem takes the following form:

$$\max_{w(x)} \quad EU_P = \int f(x|e)[x - w(x)]dx$$

$$\text{s.t.} \quad \text{(PC)} \quad EU_A = \int f(x|e)u_A[w(x)] - \alpha G[x - 2w(x)]dx - c(e) \geq U$$

$$\text{(IC)} \quad e \in \arg\max_e EU_A = \int f(x|e)u_A[w(x)] - \alpha G[x - 2w(x)]dx - c(e).$$

To solve this problem the authors rely on the first-order approach. Also setting the outside option equal to U is not – as in standard models – without loss of generality. Taking the outside option as exogenously given is like restricting the agent's reference group to the firm. If s/he is not employed in a specific firm, s/he cannot be compared with people there. Furthermore, it is assumed that the monotone likelihood ratio property holds, that is, a higher profit can serve as a signal for a higher effort choice.

Now I present the results of this model in some depth, offer a brief intuition for each of them, and compare them at each step with the standard result. Where effort is contractible and the agent is risk-neutral, the optimal contract is uniquely determined by $w' = \frac{1}{2}$, that is, the first derivative of the wage scheme specifies an equal sharing rule. This is in contrast to the standard case where there is no clear-cut prediction for the contract structure. The Principal extracts the rent with a flat payment, as agents dislike fluctuations over different levels of inequity.

If we keep effort contractible but add risk-aversion to the agent's preferences, the standard case prescribes a flat wage. With inequity aversion this no longer holds and the contract can only be shown to be increasing, as it now has to balance-off insurance against fluctuations in income and fluctuations in inequity. One can show that the slope of the incentive scheme, w', is bound between 0 and $\frac{1}{2}$.

Let us now turn to the case of interest, where effort is no longer contractible, that is the moral hazard problem. Start with considering a risk-neutral agent. Here w' is between $\frac{1}{2}$ and 1 to balance off the desire to insure against inequity and to provide strong incentives. Then adding risk-aversion to the moral hazard problem leaves us – as in the standard case – with the statement that w' is strictly increasing with profit.

The comparative statics of the latter most general case add further insights. If α, that is the agent's concern for equity, increases, the optimal contract converges to $w = \frac{1}{2}x$, that is to the equal split. In that sense, inequity aversion adds a tendency towards linear contracts. Furthermore inequity aversion is used as an additional incentive instrument. If profit x increases, the agent is not only rewarded with a higher wage payment, but also with a lower level of inequity. Thus both ways of creating utility (or reducing disutility) are used.[5]

Englmaier and Wambach's last set of results alludes to Holmström's influential sufficient statistics result. Holmström proved that optimal

contracts have to be conditional on all available informative signals (with respect to effort choice) but not to non-informative ones. The authors show that in their set-up contracts may be incomplete or overdetermined. In a situation where there is a better measure of performance than profit, that is a sufficient statistic for the effort choice, contracts should still be conditional on profit as the agent is inherently interested in profit as far as its distribution is concerned. In this sense contracts are overdetermined. If the concern for equity continues to increase, this concern for the distribution of the payoff becomes increasingly dominant. Thus the optimal contract puts less and less weight on the sufficient statistic and, in the limiting case, for extremely high values of α, disregards it altogether and is thus left optimally incomplete.

The authors relate their theoretical results to some stylized facts, such as sharecropping contracts predominantly specifying an equal split between landlord and tenant,[6] the persistence of interindustry wage differentials where more profitable firms pay higher wages to workers of the same profession,[7] and the widespread use of stock-option plans at all levels of a firm's hierarchy.[8] Englmaier and Wambach (2004) offer some additional results on the multi-agent case, and these are covered in section 5 of this chapter.

Itoh (2004) analyses a model where the agent is risk-neutral but wealth constrained. Furthermore the effort choice of the agent is not continuous but binary. His qualitative results on the structure of contracts are similar to those of Englmaier and Wambach (2003) but in addition he can show that the Principal's profit generally decreases if the agent's concern for equity increases. This result depends on the restriction that the Principal always earns more than the agent. Hence whether the Principal prefers to employ inequity averse or 'standard' agents depends on the possible profit level. If the possible profit levels are rather high, such that the Principal is better off than the agent, the Principal has to pay high wages to the agent in order to counterbalance inequity.

Dur and Glazer (2003) analyse a model where a worker envies his boss, thus neglecting the 'β-part' of Fehr and Schmidt's model. Although the workers' effort choice is continuous there are only two possible realizations of firm profits. Thus a bonus contract is optimal. Like Englmaier and Wambach, they find a violation of the sufficient statistics result. They can also show that envy increases incentive intensity but decreases the Principal's profits. They discuss several interesting applications. They suggest that envy (or more accurately a lack of it) may be a reason for less-pronounced incentives in governmental organizations. As there is no single rich Principal (or several presumably rich stockholders) towards whom the workers may feel envious, since basically the general public owns the firm, the incentive-intensifying effect present in private firms, disappears. Continuing this argument they note that progressive taxation – reducing income disparities – may in fact be efficiency-enhancing, as it dampens the adverse effects of envy. Another

application they mention is in consumer goods markets. Consumers compare themselves to the 'rich' producers of goods and are unwilling to leave too high profits to them. This affects their willingness to pay and thus restricts the producers' pricing behaviour.

4 Multiple agents: tournaments

If one now analyses situations where there is more than one agent interacting with the Principal, tournaments seem to be a natural starting setting for exploring the effects of social preferences. In a tournament, agents compete for a prize and only one of them can win. This automatically generates inequality. Tournaments are a widely studied means of providing incentives. Following the seminal contribution by Lazear and Rosen (1981), much work has been devoted to exploring the incentive properties of tournaments and to pin down situations where their use is actually optimal.

Although it is not really a tournament situation, Rey Biel's (2003) model is a good starting point to demonstrate the basic mechanism at work. He develops a deterministic model where two agents simultaneously have to make a binary effort choice. This choice is not plagued by moral hazard. The Principal can contract the agent's choice. Rey Biel uses this simple framework to highlight how the Principal can utilize the agents' inequity aversion by offering them very unequal payoffs off the desired equilibrium, and thus reduce costs. The desired equilibrium is where both exert high effort. Offering agents unequal payoffs if this outcome is not reached, inflicts disutility on them, thereby making the desired outcome, where both get the same pay, relatively more attractive. One could interpret this as a special kind of tournament where both get the prize if performance exceeds a given threshold. In this framework Rey Biel find that the agents' social preferences increase the Principal's profits. This comes as no surprise given that the Principal gets an additional instrument to generate incentives. However, the analysis neglects the agents' participation constraint, which the author justifies by arguing that the agents would also face inequity in alternative occupations.

Turning to more standard tournament settings with stochastic production and just one agent winning the prize, renders disregarding the participation constraint less innocuous than one might think, because now the agent has to be compensated upfront for the inequity inflicted on him in order to create incentives. Taking the participation constraint into account thus changes the picture quite dramatically.

Grund and Sliwka (2005) do so. They analyse a simple tournament where two agents compete for a prize. The one with the higher output wins where output is a function of effort plus an error term. They call that part of Fehr and Schmidt's model where agents suffer from being worse-off 'envy', and that part where agents suffer from being better-off, 'compassion'. For a given prize structure they show that profits unambiguously increase with an increase

in the agents' degree of envy and decrease with an increase in the agents' degree of compassion. As agents expect to feel envy if their opponent wins the tournament, they have an incentive to work harder in order to avoid this. On the other hand, if they are compassionate they are not so happy about winning. The latter effect dampens the incentive to work hard.

However, if the Principal can also choose the prize structure and wants to do so optimally, s/he has to obey the participation constraint. As in Englmaier and Wambach (2002), Grund and Sliwka (2005) assume that the outside option is exogenously given. They find that inequity aversion lowers the Principal's profits. The reason is that the agent has to be compensated upfront for the inequity that is going to be inflicted on her/him for incentive reasons, and this extra compensation outweighs the positive incentive effects. From that, Grund and Sliwka draw conclusions for a firm's optimal promotion policy. Interpreting a tournament as a competition for promotion, they compare vertical and lateral promotions. Whereas in vertical promotions the team leader is hired from within a team or group, in lateral promotions a team leader is always hired from another team. Now assuming that within a team social preferences are more pronounced, they conclude that lateral promotion schemes are preferable.

Demougin and Fluet (2003) also analyse a two-person tournament but they differ in three respects from Grund and Sliwka. First, they consider the limited liability case. Thus there can even be *ex ante* rents for the agent. Second, agents do not compare their gross payments but their rents, that is their received payments net of effort costs. And third, the Principal can invest resources in order to make the tournament more informative.

If in the initial situation the participation constraint does not bind, envy lowers the Principal's wage costs and thus increases profits. If, however, agents do not earn rents, envy and compassion both reduce profits. While the latter result is in line with Grund and Sliwka (2005) the first case is different. In the standard case, when providing only monetary incentives, the Principal leaves some rent to the agent due to limited liability. The incentives provided via the threat of inequitable outcomes are not subject to the wealth constraint and thus, for a given wealth constraint the incentive intensity is stronger. Taking into account the Principal's possibility of increasing the tournament's informational content and focussing on the envy part of inequity aversion provides additional insights. If additional precision is 'cheap' to attain, envy in fact increases profits. If however additional precision is 'expensive', profits fall.

5 Multiple agents: teams

Examining more general mechanisms than tournaments, while taking care of social preferences, the analysis of team problems becomes a more elaborate task. A first guess might be that the effect is similar to that described by

Rasmusen (1987) for risk-aversion. We can improve upon the initial situation by offering random contracts off the desired equilibrium outcome level. These random contracts have to assign the whole outcome to one player. Now agents not only face the risk of getting nothing when they shirk, but they are also likely to suffer from a large degree of inequality. Hence, as with risk-aversion, this constitutes a form of commitment to 'burn money' off the equilibrium, thus rendering a deviation less alluring. In a sense the effect is similar to the introduction of a budget-breaking Principal, who will happily keep the money if the agents have fallen short of the equilibrium effort.

But social preferences do more than just reinforce the effects of risk-aversion. Englmaier and Wambach (2004) extend their model discussed in section 3 for the case of many inequity-averse agents. They find that where for each agent an output measure is available, the optimal contract has to be conditional on each agent's individual output measure, even if the tasks are technologically independent. The reason is that by doing so the agents are offered insurance against inequitable payoffs. And in the limiting case where the agents' concern for fairness is the only important driving force, the optimal contract has a very simple structure as it is only conditional on overall output. In this way, they deliver a simple rationale for the widespread use of team-based incentives.

While in standard team-production problems the rationale for using team-based incentives (or relative performance evaluation schemes) is that this will filter out common shocks from the performance measures, Englmaier and Wambach's result is driven by the fact that agents have an inherent interest in the other agents' outcomes. Here team-based incentives are used as an insurance mechanism against very unequal outcomes.

Itoh (2004) gets results similar to those of Englmaier and Wambach (2004) but since his model is less general, he manages to pin down the contract structure rather more. In Itoh's model the two risk-neutral agents have a binary effort choice and the limited liability constraint holds. Where agents perform technologically-independent projects, he basically finds two possible contracts: an extreme team contract where all agents always get the same payment and an extreme relative performance contract which is similar to a tournament. The extreme team contract is optimal if either agent is highly inequity-averse or the project is very risky. Note that in this case the Principal's payoff is independent of the degree of inequity aversion, as agents are always paid the same. In the opposite case the extreme relative performance contract is optimal. Here the Principal generates inequity and makes use of it.

Allowing for correlated shocks to the two projects, standard theory would call for the relative performance contract. But due to a sufficient amount of inequity-aversion, the extreme team contract may remain optimal in this case. In another specification analysed by Itoh, agents do not compare their gross payments but their rents, net of effort costs. Under this assumption, he

can show that the team contract is more likely to be optimal, meaning that it is optimal for a larger set of parameter constellations.

Bartling and von Siemens (2004a) analyse a situation with deterministic team production. They require contracts to be budget-balancing and renegotiation proof. Starting from that they construct an equilibrium where the optimal contract is 'equal at the top', that is it gives an equal share to every worker if all (or all but one) agents choose high effort and assign the whole output (deterministically) to just one agent otherwise. With this mechanism they find that inequity-aversion is beneficial. However, this positive effect decreases with team size. They interpret this to be the reason why small work teams seem to perform better than larger ones. They then distinguish between worker-owned firms, that is firms with no Principal claiming residual output for himself, and firms with a Principal. They find that worker-owned firms may be inefficiently small, as agents may not want to employ an additional worker even if it were efficient to do so. They anticipate that overall surplus will be shared evenly and they may be better-off with their share in the smaller firm than in the (more efficient) larger firm. This effect is absent if there is a Principal running the firm.

In a companion paper Bartling and von Siemens (2004b) analyse a team-production setting with stochastic production for a restricted class of utility functions. There are two agents who have to make a binary effort choice, and the agents' projects are technologically independent. Here agents are not inequity-averse but only suffer from envy. Assuming inequity aversion instead may, however, invert their results. They show that envy unambiguously increases agency costs. In order to insure against the risk of suffering from envy, the Principal has to give equitable flat-wage contracts instead of incentive contracts and, as in Englmaier and Wambach (2004), team-based contracts may become optimal. In order to avoid these effects of envy, the Principal may prefer to employ only one agent although it would be efficient to employ the other too. The authors also ask whether salaries should be kept secret. Interestingly they find that keeping salaries secret is a bad idea as it takes away the chance to insure against relative income fluctuations by making one worker's pay conditional on the other workers' pay.

Masclet (2002) extends the standard team-production game with an additional stage where inequity-averse agents can punish their shirking colleagues. They will do so in order to re-establish equity. As in public goods games, described for instance by Fehr and Gächter (2000), the efficient cooperative outcome now becomes implementable. This is very close in spirit to Kandel and Lazear's (1992) model of peer pressure.

Huck and Rey Biel (2003), too, extend the standard team-production framework. They analyse a two-player situation with an exogenously-given equal sharing rule. Both agents are again inequity-averse. But here they explore what happens if agents can choose their effort sequentially. In their example they show that moving sequentially (with the less-productive agent starting)

can improve the situation because the agent that moves first can push the one that follows to a higher level of effort by choosing higher effort himself. The agent who moves later does not want to fall short of the first one's contribution. Their result is driven by their assumption that agents do not compare gross payments but payments net of effort costs.

There are two more models that provide interesting insights on the interaction of incentives and social preferences which, although not based on inequity-aversion, I discuss here.

In Rob and Zemsky's (2002) dynamic model, people are not inequity-averse but they can build up social capital. Agents can decide whether to help each other or to produce on their own. Helping is efficient but not contractible. In the repeated game, agents are more willing to help if they have received help from the other agent before, that is if social capital has been built up. A contractible performance measure is also available. Putting more weight on this contractible performance measure reduces the incentive to help, as producing more output oneself becomes relatively more lucrative. Choosing different dynamic incentive structures can give rise to 'cultural' differences across firms.

In Huck, Kübler and Weibull (2003) agents are concerned about adherence to a social norm which emerges endogenously. They restrict themselves to linear contracts and analyse the effects of such social norms in two settings. When only overall team output is observable, the social norm fosters positive externalities. If the others work more, an agent is also expected to work more in order to adhere to the social norm. This increases team output and everybody's pay. In this situation multiple equilibria exist. The authors ask whether dynamically adjusting the slope of the incentive scheme can help the Principal to select the most profitable equilibrium. If individual output is observable, relative performance schemes are utilized. If an agent now exerts more effort this has a negative effect on the relative performance of the other agents. There are now negative externalities and the norm may compress effort. The overall effect of social norms is thus unclear.

6 Conclusion

This survey has shown how incorporating social preferences in economic models can enhance our understanding of relationships in the workplace. Social preferences interact in non-trivial ways with incentives and alter the structure of optimal compensation schemes, sometimes drastically. But the research on these issues is still in its infancy.

So far the results are inconclusive with respect to the question: under what circumstances is a fair-minded workforce desirable? Related issues are the implications of social preferences for structuring work teams, the production process, the information environment or even the boundaries of the firm. These topics deserve further investigation.

Yet another interesting question is the interplay between extrinsic and intrinsic motivation and whether the provision of high-powered monetary incentives might crowd out intrinsic motivation. One could guess that these high-powered incentives change the nature of interaction and thus affect the way social preferences come into play. For further discussion of these topics see, for example, Fehr and Rockenbach (2002) or Gneezy and Rustichini (2000).

As already alluded to by Rotemberg (1994), the commitment power provided by social preferences for principals or team leaders needs investigation. Research along these lines may shed light on the determinants of good leadership and trust.

Finally, it should by now be clear that the definition of the reference group (peer group) and the definition of what actually is the job and the surplus generated from it are very important for the analysis. Those concepts, familiar to researchers in human resources, have so far received insufficient attention from economic theorists.

In conclusion, incorporating social preferences into models of agency can open the door to a fruitful dialogue between economic theorists and human resource researchers, and can prove to be a promising new avenue for research.

Notes

1 On that, see for example, Gächter and Fehr (2002).
2 This is so in the sense that at the current wage level, jobless workers would have been willing to work.
3 However, see the recent paper by Cox and Friedman (2002) who try to build a 'tractable model of reciprocity'.
4 And in the earlier version, Englmaier and Wambach (2002), as presented at the 13th World Congress of the International Economic Association in Lisbon in September 2002.
5 Mayer and Pfeiffer (2004) analyse a version of the Englmaier and Wambach model. They restrict contracts to be linear, utility exhibits constant absolute risk-aversion and the agent chooses the mean of a normal distribution. They can solve this model and confirm the findings by Englmaier and Wambach (2002).
6 Compare, for example, Bardhan (1984), and Bardhan and Rudra (1980).
7 Compare, for example, Thaler (1989) in a meta study.
8 Compare, for example, Oyer and Schaefer (2003).

References

Akerlof, G.A. (1982) 'Labour Contracts as a Partial Gift Exchange', *Quarterly Journal of Economics*, vol. 97(4) , pp. 543–69.

Akerlof, G.A. and J.L. Yellen (1988) 'Fairness and Unemployment', *American Economic Review*, vol. 78(2), pp. 44–9.

Bardhan, P. (1984) *Land, Labour and Rural Poverty* (New York: Columbia University Press).

Bardhan, P. and A. Rudra (1980) 'Terms and Conditions of Sharecropping Contracts: An Analysis of Village Survey Data in India', *Journal of Development Studies*, vol. 16(3), pp. 287–302.
Bartling, B. and F. von Siemens (2004a) 'Efficiency in Team Production with Inequity Averse Agents', Working Paper, University of Munich.
Bartling, B. and F. von Siemens (2004b) 'Inequity Aversion and Moral Hazard with Multiple Agents', Working Paper, University of Munich.
Bewley, T.F. (1999) *Why Wages Don't Fall During a Recession* (Cambridge MA: Harvard University Press).
Bolton, G.E. and A. Ockenfels (2000) 'ERC – A Theory of Equity, Reciprocity and Competition', *American Economic Review*, vol. 90(1), pp. 166–93.
Cox, J.C. and D. Friedman (2002) 'A Tractable Model of Reciprocity and Fairness', Working Paper, Learning and Experimental Economics Projects of Santa Cruz (LEEPS).
Demougin, D. and C. Fluet (2003) 'Inequity Aversion in Tournaments', Cahiers de recherche no. 0322, Centre Interuniversitaire sur le Risque, les Politiques Economiques et l'Emploi.
Dufwenberg, M. and G. Kirchsteiger (2004) 'A Theory of Sequential Reciprocity', *Games and Economic Behavior*, vol. 47(2), pp. 268–98.
Dur, R. and A. Glazer (2002) 'Optimal Incentive Contracts When Workers Envy Their Bosses', Discussion Paper, Tinbergen Institute no. 04-046/1.
Englmaier, F. and A. Wambach (2002) 'Contracts and Inequity Aversion', Economic Studies & Ifo Institute for Economic Research Working Paper no. 809, Munich.
Englmaier, F. and A. Wambach (2004) 'Contracts and Inequity Aversion', Working Paper, University of Munich.
Falk, A. and U. Fischbacher (1998) 'A Theory of Reciprocity', Institut für Empirische Wirtschaftsforschung, Working Paper no. 6, University of Zurich.
Fehr, E. and A. Falk (2002) 'Psychological Foundations of Incentives', *European Economic Review*, vol. 46, pp. 687–724.
Fehr, E. and S. Gächter (2000) 'Cooperation and Punishment in Public Goods Experiments', *American Economic Review*, vol. 90, pp. 980–94.
Fehr, E., S. Gächter and G. Kirchsteiger (1997) 'Reciprocity as a Contract Enforcement Device: Experimental Evidence', *Econometrica*, vol. 65, pp. 833–60.
Fehr, E., G. Kirchsteiger and A. Riedl (1993) 'Does Fairness Prevent Market Clearing?', *Quarterly Journal of Economics*, vol. 108, pp. 437–60.
Fehr, E. and B. Rockenbach (2002) 'Detrimental Effects of Sanctions on Human Altruism', *Nature*, vol. 422, pp. 137–40.
Fehr, E. and K.M. Schmidt (1999) 'A Theory of Fairness, Competition and Cooperation', *Quarterly Journal of Economics*, vol. 114(3), pp. 817–68.
Fehr, E. and K.M. Schmidt (2003) 'Theories of Fairness and Reciprocity – Evidence and Economic Applications', in M. Dewatripont *et al*. (eds), *Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society*, Vol. 1 (Cambridge: Cambridge University Press), pp. 208–57.
Gächter, S. and E. Fehr (2002) 'Fairness in the Labour Market? – A Survey of Experimental Results', in F. Bolle and M. Lehmann-Waffenschmidt (eds), *Surveys in Experimental Economics. Bargaining, Cooperation and Election Stock Markets* (Heidelberg: Physica Verlag), pp. 95–132.
Gneezy, U. and A. Rustichini (2000) 'Pay Enough or Don't Pay At All', *Quarterly Journal of Economics*, vol. 115(2), pp. 791–810.

Grund, C. and D. Sliwka (2005) 'Envy and Compassion in Tournaments', *Journal of Economics and Management Strategy*, vol. 14, pp. 187–207.
Holmström, B. (1979) 'Moral Hazard and Observability', *Bell Journal of Economics*, vol. 10, pp. 74–91.
Huck, S., D. Kübler and J. Weibull (2003) 'Social Norms and Economic Incentives in Firms', Working Paper, University College London.
Huck, S. and P. Rey Biel (2003) 'Inequity Aversion and the Timing of Team Production', Working Paper, University College London.
Itoh, H. (2004) 'Moral Hazard and Other-Regarding Preferences', *Japanese Economic Review*, vol. 55, pp. 18–45.
Kandel, E. and E.P. Lazear (1992) 'Peer Pressure and Partnerships', *Journal of Political Economy*, vol. 100(4), pp. 801–17.
Lazear, E.P. and S. Rosen (1981) 'Rank-Order Tournaments as Optimum Labour Contracts', *Journal of Political Economy*, vol. 89(5), pp. 841–64.
Masclet, D. (2002) 'Peer Pressure in Work Teams: The Effects of Inequity Aversion', GATE Working Paper, no. 02–15, University of Lyons.
Mayer, B. and T. Pfeiffer (2004) 'Prinzipien der Anreizgestaltung bei Risikoaversion und sozialen Präferenzen', *Zeitschrift für Betriebswirtschaft*, no. 10, pp. 1047–61.
Oyer, P. and S. Schaefer (2003) 'Why Do Some Firms Give Stock Options to All Employees?: An Empirical Examination of Alternative Theories', Stanford Business School Research Paper, no. 1772R.
Rabin, M. (1993) 'Incorporating Fairness into Game Theory and Economics', *American Economic Review*, vol. 83(5), pp. 1281–302.
Rasmusen, E. (1987) 'Moral Hazard in Risk-Averse Teams', *RAND Journal of Economics*, vol. 18(3), pp. 428–35.
Rey Biel, P. (2003) 'Inequity Aversion and Team Incentives', Working Paper, University College London.
Rob, R. and P. Zemsky (2002) 'Social Capital, Corporate Culture, and Incentive Intensity', *RAND Journal of Economics*, vol. 33(2), pp. 243–57.
Rotemberg, J. (1994) 'Human Relations in the Workplace', *Journal of Political Economy*, vol. 102(4), pp. 684–717.
Thaler, R.H. (1989) 'Anomalies: Interindustry Wage Differentials', *Journal of Economic Perspectives*, vol. 3(2), pp. 181–93.

7

Mutual Concern, Workplace Relationships and Pay Scales*

Ottorino Chillemi
University of Padua, Italy

1 Introduction

A relatively novel area of research in economic theory is how the quality of human relationships in the workplace affects reward systems and labour productivity. Akerlof's gift exchange model (1982) is a path-breaking contribution. In an attempt to explain a firm's egalitarian wage policy towards a group of its employees in a context where the average productivity in the group was also well above the standard required, Akerlof introduces both preferences for income equality among co-workers and loyalty to group norms on the part of the workers and the firm. With these assumptions he succeeds in explaining conduct that appears irrational within the standard neoclassical framework. Another interesting paper is that by Kandel and Lazear (1992), which seeks to explain how profit-sharing plans can have beneficial incentive effects. The authors focus on the role of peer pressure in curbing the incentive to free-riding inherent in such plans, and emphasize that guilt can be the only effective form of pressure when individual effort is not observable; hence the importance of empathy in motivating co-workers not to cheat on each other. In a similar vein, Rotemberg (1994) investigates whether friendly relations in the workplace can induce altruistic feelings among co-workers, thus helping to solve the free-rider problem in team production. The author discusses at length the experimental study on the productivity effects of incentive pay and labour group practices, which Mayo carried out at Hawthorne Works (Mayo, 1933). He claims that altruistic

* I am indebted to Giovanni Colombo, Gianni De Fraja, Benedetto Gui and Antonio Nicolò for insightful conversations. I am grateful to the IRC, University of Minnesota, for its hospitality. The financial support of the Italian Ministry for Universities is gratefully acknowledged.

preferences, together with strategic complementarity of individual actions, offer a sound explanation of the results.[1] As to the reasons why a person should be altruistic towards another, Rotemberg first notes that altruism may be necessary to enjoy the company of others, and then suggests that a person may materially benefit from becoming altruistic towards those he feels attracted to: 'By becoming altruistic toward you, I am led to change my behaviour. If the resulting changes in actions induce you to change your own actions in a way that benefits me, my becoming altruistic is smart' (*op cit.*: 712). Finally, in a team production setting he shows that the strategic complementarity of efforts is sufficient for altruism to be individually rational, that is, for altruism to emerge when each worker individually chooses his degree of altruism with the intent of maximizing his own material surplus. Crucial for Rotemberg's results is the fact that good-fellowship allows each worker to recognize the true attitude of his fellows, and also makes commitment possible. (For more on how people signal their moral sentiments see Frank, 1988.)

In this chapter, I build on Rotemberg's notion of trusty altruism to develop a possible explanation for the fact that firms rarely adopt pay schemes based on worker competition (see Baker, Jensen and Murphy, 1988; Gibbons, 1998). I consider a work environment with no technological interdependences in which a labour contest with a sole prize is the most profitable effort-enhancing scheme when workers are selfish. Section 2 below presents the model. Section 3 first studies the performance of a labour contest in the presence of altruism, and then discusses whether altruism can emerge endogenously. Section 4 characterizes the most profitable incentive scheme. Section 5 provides concluding comments.

2 The model

A profit-maximizing Principal has a fixed budget, normalized at 1, to spend for motivating N workers. His revenue from motivating workers is $\sum_{i=1}^{N} Ke_i$, where e_i is worker i's extra effort and K is a given positive number. Workers are risk-neutral. The maximum extra effort worker i is willing to exert to gain expected prize $p \in [0,1]$ is pv_i. We shall call pv_i worker i's valuation of prize p and $pv_i - e_i$ worker i's material surplus. We posit $v_i \in T = [v_-, v^+]$, with $v^+ > v_- \geq 0$; furthermore the viability condition $Kv^+ > 1$ holds. Valuation v_i is private information of worker i; for non-informed players each worker's valuation is a random number, drawn independently from the same distribution F, which is strictly increasing and at least twice continuously differentiable on T, with $F(v_-) = 0, \quad F(v^+) = 1$. The probability density of v_i on T is f. The following standard regularity condition holds: $f(v_i)/(1 - F(v_i))$ is strictly increasing in $v_i \in T$. Workers may be altruistic: in determining his effort, worker i takes into account both his expected material surplus S_i^p and his external benefit S_i^a, the latter consisting of the arithmetic mean of his co-workers' expected material surpluses. We limit the intensity of altruism

by assuming that workers weigh a unit of material surplus strictly more than a unit of external benefit. Since we want to use the theory of linear revelation mechanisms with symmetric agents, we specify worker *i*'s payoff as $U_i = S_i^p + \lambda S_i^a$, where $\lambda \in [0, a]$, with $a < 1$, is an altruism parameter exogenously given and equal for all workers; $\lambda = 0$ parameterises the case of selfish workers.[2]

3 The labour contest

Two different versions of the game are worthy of study. When effort is observable but not verifiable, the Principal always spends the whole budget, for any other contract would not be credible for his employees – this is the standard tournament setting. When effort is verifiable, the Principal may spend only part of the budget or assign no prize at all – for instance, s/he may require a minimum level of effort for a prize to be assigned. Hereafter we refer to the more complex case of verifiable effort – we shall briefly comment, however, on how results change if effort is unverifiable.

Let us therefore assume that, at the outset, the Principal announces that a prize will be awarded to the worker who supplies the highest effort, provided effort exceeds the minimum required level, e_s; then workers exert effort without collusion; finally the prize is assigned. Thanks to our static framework, the contest can be fruitfully analysed via an all-pay auction, in which workers bid effort and pay their submitted bids (see Baye *et al.*, 1996, for the general analogy between a rank-order tournament and an all-pay auction). Standard arguments permit the characterization of symmetric Bayes–Nash *equilibria* of the game. We can safely assume that each worker uses a pure strategy. This is represented by a piecewise continuously differentiable function, strictly increasing in the worker's valuation whenever effort is no less than e_s. Now consider worker *i*. For any given value of λ, let him/her suppose that the co-workers' effort function is $e = B(v, e_s) : T^2 \to T$. Henceforth, to economize on notation I often omit the dependence of functions on e_s; furthermore, for given e_s, $B^{-1}(e_i)$ will denote the inverse effort function. The following lemma is trivial:

Lemma 1

$$U_i = [v_i F^{N-1}(B^{-1}(e_i^*)) I_{[e_s, v^+]}(e_i^*) - e_i^*] + \frac{\lambda}{n-1} \times \sum_{k \neq i} \left[\int_{\max(e_s, B^{-1}(e_i^*))}^{v^+} v_k F^{N-2}(v_k) f(v_k) dv_k - \int_0^{v^+} B(v_k) f(v_k) dv_k \right] \quad (1)$$

where e_i^* is a best response to B and $I_{[e_s, v^+]}(e_i^*)$ equals 1 for $e_i^* \in [e_s, v^+]$ and 0 otherwise.

The first addend in (1) equals the difference between v_i, multiplied by worker i's probability of winning the prize, and e_i^*; the probability of winning is zero when $e_i^* < e_s$, while it equals the probability that all the co-workers are of a valuation lower than v_i when $e_i^* > e_s$. The second addend is the arithmetic mean of the expected material surpluses to worker i's co-workers, multiplied by the altruism parameter. A standard solving procedure allows determination of function B. It results (see Appendix) in:

$$B(v_i, e_s) = v_i F^{N-1}(v_i) - \int_{v_0}^{v_i} [F^{N-1}(v) + \lambda v F^{N-2}(v) f(v)] dv, \quad \text{if } v_0 \leq v_i \leq v^+,$$
$$B(v_i, e_s) = 0, \quad \text{if } v_i < v_0 \tag{2}$$

where v_0 is the valuation that makes worker i indifferent between exerting e_s or zero effort, so $v_0 F^{N-1}(v_0) = e_s$. The equilibrium effort function is $B^e(v_i) = B(v_i, e_s^*)$, where e_s^* is the optimal value of e_s. The following summarizes the main findings of this section.

Proposition 1 Assume work effort is verifiable. In the equilibrium of a labour contest:

(1) Effort is increasing in the valuation of the prize; if the prize is assigned, it is assigned to the worker of the highest valuation.
(2) The Principal's expected surplus is strictly decreasing in the altruism parameter.

Proof

(1) For $v_i \geq v_0$, it is $\partial B^e(v_i)/\partial v_i = (N - 1 - \lambda)\, v_i F^{N-2}(v_i) f(v_i) > 0$. Hence the winner is the worker with the highest valuation.
(2) The Principal's expected surplus is $U_0 = NK \int_{v_0}^{v^+} B(v_i) f(v_i) dv_i - [1 - F^N(v_0)]$. Now, at the optimum it is $dU_0/d\lambda = \partial U_0/\partial\lambda = -N \int_{v^*}^{v^+} \int_{v^*}^{v} z\, F(z)^{N-2} f(z) f(v) dz dv < 0$, where v^* is the equilibrium value of v_0.

Remark 1 A worker's *ex ante* expected material surplus may be either increasing or decreasing in λ. See Example 1 below.

Intuitively, our findings can be explained as follows. Assume that $\lambda = 0$ initially and then let λ increase by a small amount. Suppose the minimum required effort and effective efforts remain fixed at the values which are optimal when $\lambda = 0$. Each worker's expected marginal gain from increasing effort, which was previously zero, now becomes negative, as winning the prize implies losing the external benefit. Therefore, for given e_s, effort decreases with λ. To oppose this effect the Principal can increase e_s. However, his/her surplus decreases, because a reduction in the effort of workers has a first order effect on his/her surplus while an increase in e_s has only a second

order effect. Finally, a worker's *ex ante* expected material surplus is affected by two opposite first order effects, hence the uncertain sign of an increase in λ.

Example 1 There are two workers, valuations are independently and uniformly distributed on the unit interval, and $K=2$. Setting $F(v)=v$, $f(v)=1$ and $v^+=1$ in (2) yields $B(v)=(1-\lambda)v^2/2+(1+\lambda)v_0^2/2$. The Principal's maximizes $2\int_{v_0}^1 2B(v)dv-\left(1-v_0^2\right)$. The equilibrium value of v_0 is $v^*=(2\lambda+3)/(4+2\lambda)$. Therefore $B^e(v)=(1-\lambda)v^2/2+((1+\lambda)/2)((2\lambda+3)/(4+2\lambda))^2$, so the Principal's expected surplus is $(11+6\lambda)/(12(2+\lambda)^2)$, which is decreasing in λ. Finally, a worker's *ex ante* expected material surplus $E_{v_i}(S^p(v_i))$ equals $\int_{v^*}^1[vF(v)-B^e(v)]dv=(1+\lambda)(5+3\lambda)/(24(2+\lambda)^3)$, which is increasing for $\lambda\leq(\sqrt{7}-2)/3$ and decreasing otherwise.

Let us now briefly consider the case of unverifiable effort. In this case it is $e_s=0$ by assumption. The following is therefore evident:

Proposition 2 Assume work effort is unverifiable. Then,

(1) Statements (1) and (2) in Proposition 1 hold; and
(2) A worker's *ex ante* expected material surplus is increasing in the altruism parameter.

The labour contest with endogenous altruism

In the static setting of the preceding section, altruism was exogenously given. However, it is interesting to investigate whether, in our context, there is a sense in which altruism is compatible with selfish rationality. We now address this issue assuming that, at the outset, workers choose their altruism parameters in order to maximize material surplus. The idea that people can choose their altruism parameters may be controversial (see Rotemberg, 1994, section 1, on this point). A possible rationale for treating altruism as a choice variable is that preferences evolve during long-lasting interactions: when deciding which relationships to engage in, people take into account the consequences that will ensue from the possible development of an altruistic attitude towards specific people (here the reference is to Becker's extended utility function, whose arguments include personal and social capital, which are responsible for the individual appreciation of material benefits to change over time; see Becker, 1996). The further assumption, that a worker only takes into account his/her own material surplus when deciding which relationships to engage in, can be justified noticing that, before entering a relationship, altruism towards future co-workers is zero. Below we consider a two-stage labour contest: in the first stage, the Principal chooses e_s and the workers choose their altruism parameters; in the second stage, the degree of altruism becomes common knowledge, each worker privately learns her/his own valuation, and then the labour contest takes place.

Collective rationality

Here workers cooperatively set a common altruism parameter to maximize the expected material surplus per worker. The main result is the following:

Proposition 3 When workers choose their altruism parameter cooperatively, a strictly positive level of altruism emerges in equilibrium.

Proof First, each player's payoff function is continuous in λ and e_s, which are both defined on closed intervals, and hence the Glicksberg theorem ensures that a Nash equilibrium certainly exists (see Fudenberg and Tirole, 1991). Second, the expected material surplus per worker, $\int_{v_0}^{v^+}[vF^{N-1}(v) - B^e(v)]f(v)dv$, has a strictly positive partial derivative with respect to λ at $\lambda = 0$, and hence a selfish Nash equilibrium cannot exist.

Individual rationality

The case in which workers set their altruism parameters non-cooperatively is much more complex, and we shall be content to show that rational altruism is possible. Let us reexamine Example 1. Consider the second stage of the game, and let $e_i = b_i(\cdot;\lambda_1,\lambda_2,e_s)$ denote worker i's effort function. The inverse effort function is $\sigma_i(e) = b_i^{-1}(e)$ for $e \geq e_s$ (henceforth, when no confusion can arise, I omit parameters in writing functions). Worker i's payoff when his/her valuation is v_i and effort e – not necessarily optimal – is:

$$\left[v_i\sigma_j(e) + \lambda_i\int_{\sigma_j(e)}^{1} v\,dv\right]I_{[v_i^\circ,1]}(v_i) + \lambda_i\int_{\sigma_j(e_s)}^{1} v\,dv\,I_{[0,v_i^\circ]}(v_i) - e - \lambda_i\int_0^1 b_j(v)dv$$
$$i = 1,2,\ \ j = 1,2,\ \ j \neq i \tag{3}$$

where v_i° is such that $v_i^\circ\sigma_j(e_s) = e_s$.

The first-order condition for a maximum is:

$$\sigma_i(e)\frac{d\sigma_j(e)}{de} - 1 - \lambda_i\sigma_j(e)\frac{d\sigma_j(e)}{de} = 0, \quad \text{if } e \geq e_s,$$
$$e = 0 \qquad \text{if } e < e_s, \quad i = 1,2,\ \ j = 1,2,\ \ j \neq i \tag{4}$$

The optimal effort functions are determined by solving the system of first-order conditions (4) and the following boundary conditions,

$$\sigma_1(e_s)\sigma_2(e_s) - e_s = 0 \tag{5}$$

$$\sigma_1(e_{\max}) = \sigma_2(e_{\max}) = 1 \tag{6}$$

Equation (5) states that the marginal worker, that is one who is indifferent between effort e_s and zero effort, must get zero material surplus. Equation (6)

requires that if both workers have maximum valuation, they exert the same effort, $e_{\max}$, independently of their altruism parameter. (Remember that each worker knows the value of his/her colleague's altruism parameter.) This system has a unique solution, as is proved in the Appendix.

Now let us consider the first stage of the game. All payoff functions are continuous in λ_1, λ_2 and e_s, which are all defined on closed intervals, and hence a Nash equilibrium certainly exists. Solving for the equilibrium of the game, however, is made difficult by the fact that no closed forms can be found to the optimal effort functions. In the Appendix it is proved that a selfish equilibrium cannot exist, since when $\lambda_1 = 0$, $\lambda_2 = 0$ and $e_s = 9/16$ – this is the optimal value of e_s when $\lambda = 0$ – a worker's unilateral deviation from selfishness is profitable. We therefore state the following:

Proposition 4 When workers choose their altruism parameters non-cooperatively, individually rational altruism can emerge in equilibrium.

To understand the intuition underlying why a unilateral deviation from selfishness can be profitable, it is instructive to compute the effort functions in a simple case. Figure 7.1 refers to the case $\lambda_1 = 0.1, \lambda_2 = 0$ and $e_s = 9/16$. As was expected on intuitive grounds, worker 1 reduces his/her effort with respect to the level exerted in the case $\lambda_1 = \lambda_2 = 0$ and $e_s = 9/16$; worker 2

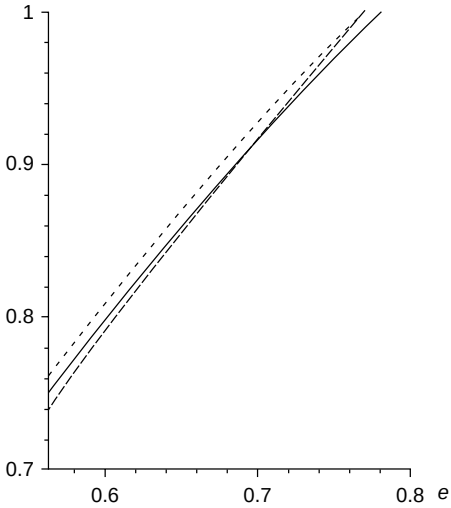


Figure 7.1 Effort functions

The solid curve is the inverse effort function for $\lambda_1 = \lambda_2 = 0$ and $e_s = 9/16$, the dotted (dashed) curve is the inverse effort function of worker 1 (2) for $\lambda_1 = 1/10$, $\lambda_2 = 0$ and $e_s = 9/16$.

increases his/her effort when s/he is of a low valuation – this is due to the rise in the probability of winning the prize – but reduces it when s/he is of a high valuation, due to the prevailing effect of the fall in the effort of the altruistic opponent. So when worker 1's effort decreases, worker 2's effort either increases or decreases, according to his/her valuation.

The general lesson of this example is that whether or not a unilateral deviation from selfishness to altruism is materially profitable depends upon the distribution of valuations. Indeed, the effort bid of a high-valuation worker is a strategic complement to the effort bid of the deviating worker, whereas the effort bid of a low-valuation worker is a strategic substitute. That is why the final outcome is uncertain.

4 The most profitable incentive scheme

The problem of finding the most profitable scheme in our setting can be addressed through direct revelation mechanisms. At the outset the Principal announces $p_i = p_i(\mathbf{v}^d) : T^N \to R_+$, and $e_i = e_i(\mathbf{v}^d) : T^N \to R$, $i \in I$, which are, respectively, worker i's prize and extra effort when $\mathbf{v}^d = (v_1^d, \ldots, v_N^d)$ is the vector of reported valuations. Then, the workers simultaneously report their valuations to the Principal, effort is exerted, and prizes are assigned according to the announced rules.

Thanks to the revelation principle we can restrict our attention to truth-telling equilibria. Let $f(\mathbf{v}) = \prod_j f(v_j)$, $d\mathbf{v} = \prod_j dv_j$, $\mathbf{v}_{-i} = (v_1, v_2, \ldots, v_{i-1}, v_{i+1}, \ldots, v_N) \in T_{-i}$, $(v_i^d, \mathbf{v}_{-i}) = (v_1, v_2, \ldots, v_{i-1}, v_i^d, v_{i+1}, \ldots, v_N)$, $f(\mathbf{v}_{-i}) = \prod_{j \neq i} f(v_j)$, and $d\mathbf{v}_{-i} = \prod_{j \neq i} dv_j$. Worker i's payoff, when s/he is of type v_i and reports v_i^d (while the other workers are telling the truth) is $U_i(v_i, v_i^d) = S_i^p(v_i, v_i) + \lambda S_i^a(v_i, v_i) = \int_{T_{-i}} [v_i p_i(v_i^d, \mathbf{v}_{-i}) - e_i(v_i^d, \mathbf{v}_{-i}) + (\lambda/(N-1)) \sum_{k \neq i} (v_k p_k(v_i^d, \mathbf{v}_{-i}) - e_k(v_i^d, \mathbf{v}_{-i}))] f(\mathbf{v}_{-i}) d\mathbf{v}_{-i}$. The Principal's expected surplus when every worker tells the truth is $U_0 = \sum_{k=1}^N \int_{T^N} [K e_k(\mathbf{v}) - p_i(\mathbf{v})] f(\mathbf{v}) d\mathbf{v}$.

The case of exogenous altruism

The characterization of the optimal scheme is not difficult. The Principal solves $Max_{\{p_i, e_i\}_{i \in I}} U_0$, s.t.: $U_i(v_i, v_i) \geq U_i(v_i, v_i^d)$, $v_i \in T$, $v_i^d \in T$, $U_i(v_i, v_i) \geq 0$, $\sum_{k=1}^N p_k(\mathbf{v}) \leq 1$, $p_i(\mathbf{v}) \geq 0$, $\mathbf{v} \in T^N$, $i \in I$, which has the same format as Myerson (1981). Let us notice a crucial difference with respect to a labour contest: in the latter S_i^p is non-negative, in the former $U_i(v_i, v_i) = S_i^p(v_i, v_i) + \lambda S_i^a(v_i, v_i)$ is non-negative. The following proposition contains the main results:

Proposition 5

(1) In the equilibrium of the optimal incentive scheme:

   (a) If the prize is assigned, it accrues to the worker of the highest valuation.

   (b) The worker of the lowest possible valuation has zero payoff.

(c) The Principal's expected surplus increases with the degree of worker altruism.

(2) When workers are selfish, a labour contest with a sole prize is optimal.

For a proof, see the Appendix. Intuitively, our finding can be explained as follows. In the presence of altruism, there is a new instrument to induce truthful reporting: workers are discouraged from lying by the fear that co-workers incur a loss. Moreover, thanks to the individual participation constraint – non-negative total payoff – the Principal can ask each non-winning worker to exert an amount of effort equal to his *ex post* external benefit weighed by the altruism parameter. An example helps to develop the intuition:

Example 2 Let $N=2, K=2, v_-=0, v^+=1, F(v)=v$. The equilibrium effort function is: $e_i(v_i, \mathbf{v}_{-i}) = (\max(v_{3-i}, v^*) - \lambda^2 v_i)/(1-\lambda^2)$ if $v_i > v_{3-i}$ and $v_i > v^*$, $e_i(v_i, \mathbf{v}_{-i}) = \lambda(v_{3-i} - (\max(v_i, v^*) - \lambda^2 v_{3-i})/(1-\lambda^2))$ if $v_i < v_{3-i}$ and $v_{3-i} > v^*$, $e_i(v_i, \mathbf{v}_{-i}) = 0$ if $v_i < v^*$ and $v_{3-i} < v^*$, where $v^* = (3+\lambda)/(4+2\lambda)$ solves $v-(1-v)/(1+\lambda) = 1/K$ (see Appendix). Notice that worker *i*'s effort decreases with his/her own valuation and increases with the co-worker's valuation. The Principal's expected surplus is $U_0 = 4\int_{v^*}^1 (v^*(v^* - \lambda^2 v)/(1-\lambda^2) + \int_{v^*}^v (u - \lambda^2 v)/(1-\lambda^2) du)dv + 4v^* \int_{v^*}^1 \lambda(v - v^*)/(1-\lambda^2) dv + 4\int_{v^*}^1 \int_z^1 \lambda(v-z)/(1-\lambda^2) dv dz - [1-(v^*)^2] = (5\lambda^2 + 16\lambda + 11)/(12(2+\lambda)^2)$, which increases with λ. A worker's *ex ante* expected material surplus is $E_{v_i} S^p(v_i, v_i) = \int_{v^*}^1 (v^*(v - v^*)/(1-\lambda^2) + + \int_{v^*}^v (v-u)/(1-\lambda^2) du)dv - v^* \lambda \int_{v^*}^1 (v - v^*)/(1-\lambda^2) dv - \lambda \int_{v^*}^1 \int_z^1 (v-z)/(1-\lambda^2) dv dz = (2\lambda^2 + 7\lambda + 5)/(24(2+\lambda)^3)$, which decreases with λ. Now, let us consider $U_i(v_i, v_i^d) = S_i^p(v_i, v_i^d) + \lambda S_i^a(v_i, v_i^d)$. At equilibrium it must be $\partial U_i(v_i, v_i^d)/\partial v_i^d|_{v_i^d = v_i} = 0$ (this is the truth-telling constraint). Now, for $v_i > v^*$ it is $\partial S_i^p/\partial v_i^d|_{v_i^d = v_i} = (\lambda/(1-\lambda^2))(1 - v_i + \lambda v_i) > 0$ and so it must be $\partial S_i^a/\partial v_i^d|_{v_i^d = v_i} < 0$.[3] This means that a worker refrains from overreporting, although overreporting increases material surplus, because his/her external benefit would decrease.

Finally, as regards the case of unverifiable effort, we notice that the only modification to the Principal's strategy with respect to the previous case is that $v^* = 0$ by assumption. Therefore, proposition 3 remains valid.

The case of endogenous altruism

We shall only comment on the findings of example 2. Since a worker's *ex ante* expected material surplus is decreasing in λ, we might conclude that, when the altruism level is chosen cooperatively, becoming altruistic is not a smart choice. However, notice that the joint surplus, that is, the sum of the Principal's and workers' surpluses, is increasing in λ.[4] Therefore, the scheme can be modified so that both the Principal and workers gain from altruism. This can be accomplished by decreasing the effort required of a worker,

independently of his valuation. (It may be recalled that we have assumed that the altruism level is decided before individual valuations are learned.) So, altruism secures workers no less material surplus than selfishness. The case in which the altruism level is chosen individually is more complex and is left to future research.

5 Summary and conclusion

We have studied how altruism among co-workers affects the performances of effort-enhancing pay schemes. First, we have proved that in a labour contest profit decreases as the level of workers' altruism increases. Second, the most profitable incentive scheme has been characterized. In sharp contrast to a labour contest, here profit increases with an increase in the level of altruism, and a firm may thus find it advantageous to foster altruistic feelings among its employees. We have also addressed the theme of endogenous altruism showing that altruism may secure workers no less material surplus than selfishness.

It is noteworthy that the optimal incentive scheme takes advantage of interdependencies among workers' payoffs due to altruism. The analysis, therefore, suggests that altruism may help explain not only the rarity of competitive pay schemes but also the wide diffusion of collective reward schemes among non-managerial personnel. An interesting theme for further research is to study what kinds of non-selfish preferences better reproduce the main features of collective reward schemes in real-life economies.

Appendix

Solving for the effort function B (section 3)

Consider v_i as a variable in equation (1). As e_i^* is optimal, it is

$$\frac{dU_i}{dv_i} = \frac{\partial U_i}{\partial v_i} = F^{N-1}(B^{-1}(e_i^*))I_{[e_s,v^+]}(e_i^*) \tag{A.1}$$

In any symmetric equilibrium, worker i will also adopt the function B, hence $e_i^* = B(v_i)$ and

$$\frac{dU_i}{dv_i} = F^{N-1}(v_i)I_{[e_s,v^+]}\left(e_i^*\right) \tag{A.2}$$

Integrating (A.2) yields another expression for worker i's payoff

$$U_i(v_i) = \int_{B^{-1}(e_s)}^{v_i} F^{N-1}(z)dz + U_i(B^{-1}(e_s)), \quad B^{-1}(e_s) \leq v_i \leq v^+ \tag{A.3}$$

Combining (1) and (A.3) yields $v_iF^{N-1}(v_i) - B(v_i) + \lambda \int_{v_i}^{v^+} vf(v)F^{N-2}(v)dv - \lambda \int_0^{v^+} B(v)f(v)dv = \int_{B^{-1}(e_s)}^{v_i} F^{N-1}(z)dz + \lambda \int_{B^{-1}(e_s)}^{v^+} vf(v)F^{N-2}(v)dv - \lambda \int_0^{v^+} B(v)f(v)dv$, hence we have $B(v_i) = v_iF^{N-1}(v_i) - \int_{B^{-1}(e_s)}^{v_i} F^{N-1}(z)dz - \lambda \int_{B^{-1}(e_s)}^{v_i} vf(v)F^{N-2}(v)dv$.

Noticing that $B^{-1}(e_s)$ is the value of v_i which makes a worker indifferent between e_s and zero effort yields the effort function in the text.

Proof of system (5)–(6) having a unique solution (section 3)

Suppose $e_{\max} \in [0, 1]$. A unique solution to (4) exists through $\sigma_1(e_{\max}) = \sigma_2(e_{\max}) = 1$, since in a neighbourhood of this point, both $1/(\sigma_1 - \lambda_1\sigma_2)$ and $1/(\sigma_2 - \lambda_2\sigma_1)$ are continuously differentiable and so locally Lipschitz. Next, the solution attains the locus defined by (5). In fact, let us suppose, for the sake of argument, that there is $e' \in (0, e_{\max})$ such that $\sigma_2(e') - \lambda_2\sigma_1(e') \leq 0$ and $\sigma_1'(e')$ unbounded. This implies $d\sigma_2/d\sigma_1|_{e=e'} \leq 0$, since $d\sigma_1/d\sigma_2 = (\sigma_1 - \lambda_2\sigma_2)/(\sigma_2 - \lambda_1\sigma_1)$. Hence it would be $d\sigma_2/d\sigma_1 < \lambda_2$ for $e' + \varepsilon, \varepsilon > 0$ small enough, which means that no trajectory could reach $\sigma_2 = \sigma_1 = 1$, a contradiction. We conclude that the solution can never cross the line $\sigma_2 - \lambda_2\sigma_1 = 0$, hence σ_1' is always bounded and strictly positive. A similar argument holds for σ_2'. To complete the proof we only have to check that $e_{\max} \in [0, 1]$ and is unique. Summing all equations in (4) yields, upon an integration, $2\sigma_2\sigma_1 - \lambda_1\sigma_2^2 - \lambda_2\sigma_1^2 - 4e = c$, where c is a constant to be determined. (4) permits determination of $c = -\lambda_1\sigma_2^2(e_s) - \lambda_2\sigma_1^2(e_s) - 2e_s$. Finally (6) implies $e_{\max} = \frac{1}{4}\{2 - \lambda_1[1 - \sigma(e_s)] - \lambda_2[1 - e_s/\sigma(e_s)] + 2e_s\} = (2 - \lambda_1 - \lambda_2 - c)/4 \in [0, 1]$, for all $e_s \in [0, 1]$.

Proof of Proposition 3

We prove that worker 1 finds it materially advantageous to unilaterally deviate from $\lambda_1 = \lambda_2 = 0, e_s = 9/16$. The thesis is true if at $\lambda_1 = \lambda_2 = 0, e_s = 9/16$ the following derivative is positive: $(\partial/\partial\lambda_1)\int_{\sigma_1(e_s)}^{1}(\nu_1\sigma_2(b_1(\nu_1)) - b_1(\nu_1))d\nu_1$, which equals $\int_{0.75}^{1}\nu_1(\partial\sigma_2(b_1(\nu_1))/\partial\lambda_1)d\nu_1$. To find $\partial\sigma_2/\partial\lambda_1$, we differentiate (4) with respect to λ_1. Using the fact that $\sigma_1(e; 0, 0, 9/16) = \sigma_2(e; 0, 0, 9/16) = \sqrt{2\,(e - 9/32)}$, and setting $g = e - 9/32$, we have $\partial\sigma_1(g)/\partial\lambda_1 + (\partial^2\sigma_2(g)/\partial g\partial\lambda_1)2g - \sqrt{2g} = 0, \partial\sigma_2(g)/\partial\lambda_1 + 2g(\partial^2\sigma_1(g)/\partial g\partial\lambda_1) = 0$. The solution to the above system is:

$$
\begin{aligned}
\frac{\partial\sigma_1(g)}{\partial\lambda} &= \frac{1}{4}\sqrt{2}\left(\cosh\left(\frac{1}{2}\ln g\right)\right)g - \frac{1}{4}\sqrt{2}\cosh\left(\frac{1}{2}\ln g\right)\ln g \\
&\quad - \frac{1}{4}\sqrt{2}\left(\sinh\left(\frac{1}{2}\ln g\right)\right)g - \frac{1}{4}\sqrt{2}\sinh\left(\frac{1}{2}\ln g\right)\ln g \\
&\quad + \left(\cosh\left(\frac{1}{2}\ln g\right)\right)C_1 + \left(\sinh\left(\frac{1}{2}\ln g\right)\right)C_2, \\
\frac{\partial\sigma_2(g)}{\partial\lambda} &= \frac{1}{4}\sqrt{2}\left(\sinh\left(\frac{1}{2}\ln g\right)\right)g - \frac{1}{4}\sqrt{2}\left(\cosh\left(\frac{1}{2}\ln g\right)\right)g \\
&\quad + \frac{1}{4}\sqrt{2}\sinh\left(\frac{1}{2}\ln g\right)\ln g + \frac{1}{2}\sqrt{2}\cosh\left(\frac{1}{2}\ln g\right) \\
&\quad + \frac{1}{4}\sqrt{2}\cosh\left(\frac{1}{2}\ln g\right)\ln g + \frac{1}{2}\sqrt{2}\sinh\left(\frac{1}{2}\ln g\right) \\
&\quad - \left(\sinh\left(\frac{1}{2}\ln g\right)\right)C_1 - \left(\cosh\left(\frac{1}{2}\ln g\right)\right)C_2.
\end{aligned}
$$

We can get boundary conditions in the following way. First, differentiation of (6) with respect to λ_1 yields $\partial\sigma_1(e_{\max})/\partial\lambda_1 + (\partial\sigma_1(e_{\max})/\partial e_{\max})(de_{\max}/d\lambda_1) = 0$, $\partial\sigma_2(e_{\max})/\partial\lambda_1 + (\partial\sigma_2(e_{\max})/\partial e_{\max})(de_{\max}/d\lambda_1) = 0$. These equations imply $\partial\sigma_1(e_{\max})/\partial\lambda_1 = \partial\sigma_2(e_{\max})/\partial\lambda_1$ for $\lambda_1 = \lambda_2 = 0$. A second boundary condition can be derived by observing that $b_1(\nu_1)$ is decreasing in λ_1, for the reason explained in the comment to

proposition 1. Therefore, when λ_1 increases by a small amount starting from zero, some workers who, for $\lambda_1 = 0$, bid zero effort, will bid a positive level of effort, thanks to the increase in their probability of winning the prize. This implies that, at $\lambda_1 = \lambda_2 = 0$, $\partial\sigma_1(e_s)/\partial\lambda_1$ is opposite in sign to $\partial\sigma_2(e_s)/\partial\lambda_1$.

Now, let us set $C_1 = C_2$, so the two derivatives are opposite in sign at $g = 0$, as the second condition requires, and $C_1 = C_2 = -0.245\,05$, so the two derivatives are equal in $g = 1/2$, as the first condition requires (notice that $B^e(v) = v^2/2 + 9/32$, so $e_{\max} = 1/2 + 9/32$). Finally, we get $\int_{9/32}^{1/2} 2g(\partial\sigma_2(g)/\partial\lambda_1)dg = 2.9706 \times 10^{-2} > 0$, and hence we conclude that worker 1 certainly benefits from a unilateral deviation to altruism.

Proof of Proposition 5

Following Myerson (1981), it is possible to show that, in any feasible mechanism, it is $U_0/K = -\sum_{i\in I} U_i(v_-, v_-)/(1+\lambda) + \int_{T^N} \sum_{i\in I}(v_i - (1/K) - (1/(1+\lambda))(1 - F(v_i))/f(v_i))p_i(\mathbf{v})f(\mathbf{v})d\mathbf{v}$. U_0 is maximized by setting $U_i(v_-, v_-)$ equal to zero for all i, and $p_i(v_i, \mathbf{v}_{-i})$ equal to one (zero) whenever $c(v_i) = v_i - (1/(1+\lambda))(1 - F(v_i))/f(v_i) > 1/K$. Under our assumptions, $c(v_i)$ is a monotonic strictly increasing function of v_i. For any vector $\mathbf{v}_{-i}$, consider $z(\mathbf{v}_{-i}) = \inf\{s: c(s) \geq 1/K \text{ and } c(s) \geq c(v_j) \text{ for all } j \neq i\}$. Then, in equilibrium $p_i(s_i, \mathbf{v}_{-i}) = 1$ if $s_i \geq z(\mathbf{v}_{-i})$, and $p_i(s_i, \mathbf{v}_{-i}) = 0$ if $s_i < z(\mathbf{v}_{-i})$, and hence $\int_{v_-}^{v_i} p_i(s_i, \mathbf{v}_{-i})ds_i$ equals $v_i - z(\mathbf{v}_{-i})$ if $v_i \geq z(\mathbf{v}_{-i})$ and equals 0 otherwise. Thus, we have proved parts 1(a) and 1(b) of the proposition.

As regards part 1(c), it is immediately seen that U_0 is increasing with λ. Finally, part 2 is a consequence of the Revenue Equivalence Theorem for auctions, which holds when $\lambda = 0$ (see for instance Myerson, 1981).

The characterization of the effort functions is as follows. By lemma 2 in Myerson (1981) we can write $U_i(v_-, v_-) = \int_{T_{-i}}[\lambda/(N-1)\sum_{k\neq i}(v_k p_k(v_i, \mathbf{v}_{-i}) - x_k(v_i, \mathbf{v}_{-i})) + v_i p_i(v_i, \mathbf{v}_{-i}) - x_i(v_i, \mathbf{v}_{-i}) - \int_0^{v_i} p_i(s_i, \mathbf{v}_{-i})ds_i]f(\mathbf{v}_{-i})d\mathbf{v}_{-i}$, $v_i \in T$, $i \in I$. Now, we set the expression in braces at zero to have $U_i(v_-, v_-) = 0$ as required in equilibrium. For doing this, we recall that $\int_{v_-}^{v_i} p_i(s_i, \mathbf{v}_{-i})ds_i$ equals $v_i - z(\mathbf{v}_{-i})$ if $v_i \geq z(\mathbf{v}_{-i})$, and equals 0 if $v_i < z(\mathbf{v}_{-i})$. Therefore, let us fix state $\mathbf{v} = (v_i, \mathbf{v}_{-i})$, and suppose a winner exists. Let $w \in I$ denote the winner. Then, for each state $\mathbf{v}$ in which a winner exists, $x_j(\mathbf{v})$, $j = 1, 2, \ldots, N$, solve the system of N linear equations:

$$\frac{\lambda}{N-1}\sum_{k\neq w} x_k(v_w, \mathbf{v}_{-w}) + x_w(v_w, \mathbf{v}_{-w}) = z(\mathbf{v}_{-w}),$$

$$\frac{\lambda}{N-1}\left(v_w - \sum_{k\neq j} x_k(v_w, \mathbf{v}_{-w})\right) - x_j(v_w, \mathbf{v}_{-w}) = 0, \quad j \in I - \{w\}.$$

Finally, in those states in which the prize is not assigned, all efforts are zero.

Notes

1 In this context two actions are strategic complements if a worker finds it profitable to exert more effort in response to an increase in a co-worker's effort.

2 This notion of altruism implies that a worker cares only about the first moment of the distribution of material surplus among co-workers, so the weight s/he gives to a co-worker's material surplus decreases as the number of co-workers increases.

3 It is $S_i^p(v_i, v_i^d) = v_i \int_0^{v_i^d} du - \int_0^{v^*} (v^* - \lambda^2 v_i^d)/(1-\lambda^2) du - \int_{v^*}^{v_i^d} (u - \lambda^2 v_i^d)/(1-\lambda^2) du - \int_{v_i^d}^1 \lambda(u - ((v_i^d - \lambda^2 u)/(1-\lambda^2))) du$.

4 It is $U_0 + 2E_{v_i} S_i^p(v_i, v_i) = (5\lambda^2 + 16\lambda + 11)/(12(2+\lambda)^2) + 2((2\lambda^2 + 7\lambda + 5)/(12(2+\lambda)^3)) = (28\lambda^2 + 5\lambda^3 + 50\lambda + 27)/(12(2+\lambda)^3)$.

References

Akerlof, G.A. (1982) 'Labour Contracts as a Partial Gift Exchange', *Quarterly Journal of Economics*, vol. 97(4), pp. 543–69.

Baker, G.P., M.C. Jensen and K.J. Murphy (1988) 'Compensation and Incentives: Practice vs. Theory', *Journal of Finance,* vol. 43(3), pp. 593–615.

Baye, M.R., D. Kovenock and C.G. de Vries (1996) 'The All-Pay Auction with Complete Information', *Economic Theory*, vol. 8(2), pp. 291–305.

Becker, G.S. (1996) *Accounting for Tastes* (Cambridge, MA and London: Harvard University Press).

Fudenberg, J. and J. Tirole (1991) *Game Theory* (Cambridge, MA: MIT Press).

Frank, R. (1988) *Passions within Reason: The Strategic Role of the Emotions* (New York and London: Penguin).

Gibbons, R. (1998) 'Incentives in Organizations', *Journal of Economic Perspectives*, vol. 12(4), pp. 115–32.

Kandel, E. and E.P. Lazear (1992) 'Peer Pressure and Partnerships', *Journal of Political Economy*, vol. 100(4), pp. 801–17.

Mayo, E.G. (1933) *The Human Problems of an Industrial Civilization* (New York: Macmillan now Palgrave).

Myerson, R.B. (1981) 'Optimal Auction Design', *Mathematics of Operations Research*, vol. 6(1), pp. 58–73.

Rotemberg, J. (1994) 'Human Relations in the Workplace', *Journal of Political Economy,* vol. 102(4), pp. 684–717.

Part II
Laboratory and Field Experiments

This page intentionally left blank

8

Expectations and the Effects of Money Illusion*

Ernst Fehr
Institute for Empirical Research in Economics, University of Zurich, Switzerland
and
Jean-Robert Tyran
Institute of Economics, University of Copenhagen, Denmark

1 Introduction

While the debate on how economic agents form expectations and how these expectations should be modelled has been key to modern macroeconomics, money illusion has been an anathema to macroeconomists until recently. The rational expectations revolution in the 1970s thoroughly banned the study of money illusion from economists' research agendas. Rational individuals do not exhibit illusions and because, by assumption, people behave rationally, there is nothing to study. Money illusion was a concept to be mentioned in courses on the history of economic thought but not a part of actual research endeavours. In fact, a reliable method for getting leading journals to reject theory papers was to propagate that money illusion affected individual behaviour.

There is an intuitively powerful argument supporting the view that money illusion is irrelevant for economics: this states that since people will suffer economically from their illusion they have a strong incentive to make illusion-free decisions. Therefore, people will eventually learn to make illusion-free decisions, implying that money illusion has little or no impact on aggregate outcomes. The purpose of this chapter is to show that this argument is seriously misleading because it neglects the indirect effects of money illusion in a strategic environment, where agents have to form expectations (including higher-order expectations; that is, expectations about expectations of others, and so on) to make optimal decisions.

* Financial support by the National Centre of Competence in Research on 'Financial Valuation and Risk Management' is gratefully acknowledged. The national centres in research are managed by the Swiss National Science Foundation on behalf of the federal authorities.

Such expectations are exceedingly difficult to form, and may be shaped by money illusion.

We show experimentally that even if money illusion only distorts individual decisions slightly, it can have important aggregate-level effects because money illusion can shape expectations. We analyse two types of aggregate-level effects. First, we show that money illusion is a cause of nominal inertia after an anticipated monetary shock in an economy with a unique equilibrium. Second, we show that money illusion can even have permanent effects by coordinating individuals on inferior equilibria. We begin our discussion by explaining that experimental methods are useful in investigating money illusion as a cause of nominal inertia because experimental methods provide insights that cannot be gained with other empirical approaches. We proceed by presenting an experimental design in which money illusion causes nominal inertia even if money illusion is almost absent at the individual level. The importance of expectation formation is demonstrated by comparing an experimental treatment in which experimental subjects play against other subjects with one in which subjects play against computerized (simulated) agents. Since subjects know that they play against computers that are programmed to have perfect foresight, they do not have to form expectations. Furthermore, we identify strategic complementarity as a key element of nominal inertia by showing that the standard rational expectations theory provides very accurate predictions of aggregate-level behaviour in the absence of strategic complementarity.

The second type of aggregate-level effects of money illusion which we analyse concerns permanent effects. While the experimental designs for the study of nominal inertia had a unique equilibrium to which nominal prices eventually converge in all cases, the design for the examination of permanent effects has *multiple* equilibria. We experimentally show that money illusion can have powerful permanent effects in an environment with multiple Pareto-ranked equilibria (arising from a locally extreme degree of strategic complementarity). These permanent effects arise because money illusion induces subjects to coordinate on inferior equilibria. Once individuals attain a bad equilibrium, they are locked out so that they experience permanent economic losses relative to the efficient equilibrium. We again compare this behaviour to a treatment in which subjects play against computerized agents with perfect foresight to isolate for the role of expectations. In this case, we find that individual-level money illusion initially causes some individuals to make non-efficient choices, but behaviour eventually converges to the efficient equilibrium in most cases. Thus, even if individual-level money illusion is only a temporary phenomenon in a non-strategic setting, it can cause permanent real effects in a strategic setting by coordinating people on inefficient equilibria.

We proceed as follows. Section 2 reports evidence from questionnaire studies suggesting that money illusion is an important phenomenon at the

individual level. Section 3 explains why economists have been interested in nominal inertia and discusses the particular strengths of the experimental approach. We also explain why expectations can magnify the effects of money illusion. Section 4 discusses experimental studies investigating money illusion as a cause of nominal inertia. Section 5 presents an experimental design for investigating the permanent effects of money illusion and reports the main findings. Section 6 concludes.

2 Money illusion at the individual level

Various authors have used the term 'money illusion' in different manners, although the foundation behind the term seems to be rather similar (see Howitt, 1989). The basic intuition says that if the *real* incentive structure, that is, the *objective* situation an individual faces, remains unchanged, the *real* decisions of an illusion-free individual also remain constant. This intuition is built on two crucial assumptions: first, the objective function of the individual does not depend on nominal but only on real magnitudes. Second, people perceive that purely nominal changes do not affect their opportunity set. For example, people have to understand that an equiproportionate change in all nominal magnitudes leaves the real constraints unaffected. Some economists suspected that these assumptions do not always hold. For example, Irving Fisher (1928) was convinced that ordinary people, in general, fail 'to perceive that the dollar, or any other unit of money expands or shrinks in value' after a monetary shock.

However, whether people are indeed able to 'pierce the veil of money' is an empirical question. Shafir, Diamond and Tversky (henceforth SDT, 1997) conducted questionnaire studies indicating that frequently one or both preconditions for the absence of money illusion are violated. Their results suggest that nominal values affect both people's preferences as well as their perceptions of the constraints. Moreover, many people not only seem to be prone to money illusion; they also expect other people's preferences and decisions to be affected by money illusion. Problem 1 of SDT's questionnaire study neatly illustrates these claims. SDT presented the following hypothetical scenario to two groups of respondents:

> Consider two individuals, Ann and Barbara, who graduated from the same college a year apart. Upon graduation, both took similar jobs with publishing firms. Ann started with a yearly salary of $30,000. During her first year on the job there was no inflation, and in her second year Ann received a 2% ($600) raise in salary. Barbara also started with a yearly salary of $30,000. During her first year on the job there was 4% inflation, and in her second year Barbara received a 5% ($1,500) raise in salary.

Respondents of group 1 were then asked the happiness question: *'As Ann and Barbara entered their second year on the job, who do you think was happier?'* 36 per cent thought that Ann was happier while 64 per cent believed that Barbara was happier. This indicates that most subjects believed that preferences are affected by nominal variables because in real terms, of course, Ann does better.[1] Respondents of group 2 were asked the following question: *'As they entered their second year on the job, each received a job offer from another firm. Who do you think was more likely to leave the present position for another job?'* In line with the response to the happiness question, 65 per cent believed that Ann, who is doing better in economic terms, is more likely to leave the present job. Thus, a majority believed that other people's decisions are affected by money illusion.

Since the absence of money illusion means that purely nominal changes do not affect an individual's preferences, perceptions nor, hence, choices of real magnitudes, it is natural to view *money illusion as a framing effect*. From this viewpoint, an individual exhibits money illusion if the preferences or the perception of the constraints and the associated decisions depend on whether the same environment is represented in nominal or real terms. SDT's analysis is based on a large body of research in cognitive psychology that shows that alternative representations of the same situation may well lead to systematically different responses (Tversky and Kahneman, 1981; 1986). Representation effects seem to arise because people tend to adopt the particular frame that is presented and evaluate the options within this frame. Because some options loom larger in one representation than in another, an alternative framing of the same option can give rise to different choices.

SDT argue that people tend to have multiple representations but that the nominal representation is often *simpler and more salient*. They suggest that people are generally aware of the difference between nominal and real values, but because money is a salient and *natural unit*, people often think of transactions predominantly in nominal terms.

Economists tend to question the relevance of results from questionnaire studies on two grounds. First, they suspect that there may be a considerable difference between what people say they would do in a hypothetical scenario and what they actually do when subject to economic incentives. Second, it is not sufficient to show that money illusion prevails at the individual level to conclude that money illusion will be of any importance at the aggregate level from an economic viewpoint. For example, the individual-level effects of money illusion may cancel out with interaction and may therefore be irrelevant at the aggregate level. Experimental methods enable the researcher to address these objections. The interaction between economic agents which are exposed to economic incentives can be studied in experimental investigations.

3 Expectations and the aggregate effects of money illusion

In this section, we explain how expectations formation can yield large indirect effects of money illusion even if the direct (that is, individual-level) effects of money illusion are small. If these indirect effects are important, money illusion has important aggregate-level effects. Below, we discuss two such effects: long-run effects arising from permanent miscoordination on inferior equilibria, and short-run effects arising from nominal inertia.

Nominal inertia refers to a tendency of *nominal* prices and wages to adjust slowly to nominal shocks. One of the reasons why economists have been interested in nominal inertia ever since the writings of David Hume (1752) is that nominal inertia implies monetary non-neutrality, meaning that nominal inertia implies that changes in monetary policy affect real macroeconomic variables like output or employment. In principle, money illusion could provide an explanation for the inertia of nominal prices and wages; such explanations were routinely invoked before the advent of the rational expectations revolution of the 1970s.[2] However, the notion of money illusion seems to have been thoroughly discredited in mainstream economics in the meantime. Tobin (1972), for example, described the negative attitude of most economic theorists towards money illusion as follows: 'An economic theorist can, of course, commit no greater crime than to assume money illusion.' The reason for this negative attitude is that money illusion contradicts basic rationality assumptions and does not fit nicely into the equilibrium mould of economics. As a consequence, economists have sought explanations of nominal inertia which are based on the assumption of fully rational agents holding rational expectations. For example, factors like informational frictions (Lucas, 1972), staggering of contracts (for example, Fischer, 1977; Taylor, 1979) and costs of price adjustment (Mankiw, 1985) have been invoked to explain nominal inertia in a fully rational framework.

The inertia of nominal prices and wages has been deemed an important phenomenon (see, for example, Akerlof, Dickens and Perry, 1996; Kahn, 1997). However, despite the vast amount of empirical and theoretical literature on nominal inertia, very little is known about its *causes*. One of the reasons for this lack of knowledge is that the empirical research strategies applied to date were inept for isolating the causes of nominal inertia. For example, Alan Blinder and his colleagues (1998: 3) ask: 'Why are wages and prices so "sticky"? The abject failure of standard research methodology to make headway on this critical issue in the micro-foundations of macroeconomics motivated the unorthodox approach of the present study.' The 'unorthodox' approach chosen by Blinder and his colleagues is to ask managers about how and why they change prices, while our unorthodox approach is to conduct economic laboratory experiments. In the next section, we argue that experimental methods allow a new examination of this old and important issue. In section 4, we develop an experimental framework for

investigating whether money illusion causes nominal inertia. We investigate the adjustment of nominal prices after an anticipated monetary shock in an environment in which firms face no exogenous obstacles to price adjustment whatsoever. Therefore, none of the rationality-based explanations for nominal inertia mentioned above apply. As a consequence, our investigation does not intend to question the potential relevance of these rationality-based explanations. However, our results do demonstrate the importance of money illusion, expectation formation and strategic complementarity in understanding the causes of nominal inertia. The results suggest, in particular, that money illusion has been dismissed prematurely as a candidate for the explanation of nominal inertia.

Why use experimental methods to investigate nominal inertia?

In laboratory experiments, we observe the behaviour of real people who are exposed to real economic incentives in a controlled environment. In what respect do experimental investigations have advantages over empirical investigations with field data? An obvious first advantage consists of the *correct measurement* of endogenous variables like prices and real economic activity. In contrast, the conclusions drawn from field studies investigating the real effects of monetary shocks appear to be extremely non-robust with respect to measurement problems (see for example, Belongia, 1996). Second, data that is crucial for many economic theories can be gathered in the laboratory but cannot be directly observed in the field. *Expectation data* are especially valuable in our context. The third and most important advantage of the experimental method is *control over the environment and the information conditions*. The ability to control the environment has several implications. For example, truly *exogenous* monetary shocks can be implemented in the laboratory. In contrast, macroeconomic field studies are plagued by notorious causality problems. In the laboratory, the theoretical equilibrium values of the economy under study are known. Therefore, the observing experimentalist can distinguish between equilibrium and out-of-equilibrium realizations of endogenous variables. This is a crucial advantage, since nominal inertia is a disequilibrium phenomenon. In addition, we control information conditions, that is, we control what economic agents know about their economic environment and what they know about the information available to other agents. As explained in section 4, this allows us to implement an anticipated monetary shock.

Finally, *causal relations* can be established in an experiment through controlled *ceteris paribus* variations in the decision environment. The causes of nominal inertia can be isolated by changing only one aspect of the environment and by comparing nominal price adjustment in the respective treatments. Our main objective is to investigate whether money illusion is a cause of nominal inertia. As explained above, money illusion implies that behaviour depends on whether the same objective situation is framed

in nominal or in real terms. A particularly transparent example of money illusion prevails if people behave differently when they receive payoff information in real or in nominal terms. Unfortunately, business life does not seem to provide examples where the same objective situation is sometimes represented in nominal terms and sometimes in real terms. In fact, almost all business transactions involve nominal payoff information. Therefore, a major advantage in the experimental approach to the causes of nominal inertia is that the 'frame' is under the experimenter's control. In particular, we implemented a treatment condition in which payoffs were represented in nominal terms and a control condition in which payoffs were represented in real terms.

The above arguments suggest that experimental methods can be very useful for the examination of nominal inertia (see Duffy, 1998, for a survey of experiments in monetary economics). We also would like to stress, however, that these methods are not a substitute for the analysis of field data. Laboratory experiments, in our view, complement the standard econometric techniques of the analysis of field data. Both methods should be used to increase our knowledge. Yet, the marginal return from experimental methods is likely to be high due to the relative lack of these investigations in the past.

Money illusion and expectations in a strategic environment

This section explains that money illusion can have both direct and indirect effects. The direct effects arise from individual optimization errors resulting from a confusion of nominal and real values. The indirect effects occur in a strategic environment and are shaped by expectations. To understand these indirect effects, three elements need to be introduced. First, agents are heterogeneous with respect to rationality. A large body of experimental evidence shows that the assumption of rationality does not hold equally well for all agents in a variety of contexts (see Camerer, 2003, for a survey). In our context, this simply means that some agents are prone to individual-level money illusion while others are not. The second element is strategic complementarity which essentially induces an incentive to 'follow the crowd'. In principle, these two elements are sufficient to explain indirect effects of money illusion. However, some theoretical models using these two elements make the unrealistic assumption of perfect foresight among rational agents. We must therefore introduce a third element illustrating that expectations are uncertain and biased by money illusion, which we will explain in detail below.

Strategic complementarity means that if other agents change the value of their action variable (for example, their prices), it is optimal for a rational agent to change the value of his action variable (for example, his price) in the same direction. It has been argued that strategic complementarity is an important characteristic of macroeconomic relations (Cooper and Haltiwanger, 1996), and it certainly is a natural property of (monopolistic)

price competition. Technically speaking, strategic complementarity implies a positive slope of the reaction function. The intuition behind the concept is that rational agents have an incentive to 'follow the crowd'. Yet, to be able 'to follow the crowd', agents have to predict where it is going, so to speak. More generally, agents have to form expectations to make optimizing decisions in a strategic environment.

In a situation with strategic complementarity, heterogeneity of agents can multiply the effects of individual-level bounded rationality. For example, Haltiwanger and Waldman (1985, 1989) show theoretically that under strategic complementarity, a *small* group of non-rational price setters can have *large* effects on the adjustment process to equilibrium because they induce rational agents to hold non-equilibrium expectations and to choose non-equilibrium actions. The intuition behind this theoretical result is that the rational agents partially imitate the behaviour of the non-rational agents because of strategic complementarity and thereby multiply the effects of the latter on the aggregate price level. However, this model relaxes the rationality assumption in a specific manner. On the one hand, there are non-rational agents who do not optimally adjust their nominal prices to the monetary shock. On the other hand, the rational agents in fact are assumed to have perfect foresight. These agents surmise correctly how the presence of non-rational agents affects the economy. Note that forming higher-order expectations is a very difficult task; it requires the ability to predict other peoples' predictions about what everybody else does, for example. Therefore, the rationality requirements are unrealistically high in this model.[3]

Suppose that human decision-makers are rational but do not have perfect foresight. That is, suppose decision-makers are able to choose a rational action given that they know what everybody else does, but unable to form correct expectations. In particular, expectations are assumed to be uncertain and biased by money illusion in a strategic environment. In our context, this means that decision-makers assume with positive probability that some others suffer from money illusion (or assume that others expect again others to suffer from money illusion, and so on). As a result, they assume that this group will not fully adjust to the shock. This uncertain and biased expectation, in turn, also induces the rational agents to adjust only imperfectly to a shock. An interesting aspect of this idea is that a mere biased belief about others' money illusion may cause nominal inertia even if no single individual is in fact prone to money illusion. That is, given strategic complementarity, money illusion may cause pronounced nominal inertia even if individual-level money illusion is very small or non-existent.

The illusion-based account of nominal inertia provided above is structurally identical to an 'informational friction' theory which has recently met with renewed interest. For example, Woodford (2003) as well as Mankiw and Reis (2002) impose a constraint on the information that people use when forming expectations, and, as a consequence, only a share of agents is fully

informed about shocks at any time. To justify these informational constraints, the authors cite work by Sims (2003) on 'rational inattention'.[4] However, the apparent similarity of assuming limited information processing capacity and limited rationality does not seem to be discussed in the literature,[5] nor is it apparently essential to this line of research (instead, the main aim of this research is to investigate the implications of individual information constraints for optimal monetary policy; see Adam, 2003, or Ball, Mankiw and Reis, 2003).

4 Experiments on money illusion as a cause of nominal inertia

In this section we first provide a brief description of the basic design, followed by the hypotheses and the main results of an experimental study on money illusion as a cause of nominal inertia after a negative monetary shock. We then briefly report on an experiment with a positive shock, and investigate how strategic properties affect nominal inertia.

Design

In our experiment, $n = 4$ subjects are in the role of firms and simultaneously choose nominal prices in T consecutive periods. The firms are free to change nominal prices at no (menu) cost in any period. Price competition is characterized by strategic complementarity; that is, if an agent expects his competitors to increase their prices, it is optimal for him to do so as well. After the simultaneous pricing decision, each firm receives information about the aggregate price level (resulting from other firms' choices) and the own payoff. Firms take their decisions in a fully stationary environment, that is, there is no exogenous uncertainty whatsoever. The experiment has a pre-shock and a post-shock phase with a length of $T/2$ periods each. The pre-shock phase mainly serves the purpose of equilibrating the system. The post-shock phase serves to observe how nominal prices adjust to the monetary shock in various treatment conditions (for details see Fehr and Tyran, 2001).

The real payoff of player i, π_i, depends on his own nominal price p_i, on the *average* price competing firms choose P_{-i} (that is, the price level, excluding the choice of i), and on a nominal shift variable (the quantity of money M) in the following way: $\pi_i = \pi_i(p_i/P_{-i}, M/P_{-i})$. Thus, the real payoff remains unchanged if all prices and M change by the same percentage. Subjects receive information about payoffs in payoff tables (payoff matrices). This is possible because the payoff depends only on p_i and P_{-i} for a given level of M. The payoff table shows the nominal or the real payoffs of a player for all feasible combinations of p_i and P_{-i}. The treatment condition determines whether the payoff table shows the nominal or the real payoff (see below). All players are fully informed about their own payoff table and those of the other $n - 1$ players in the group.

The monetary shock is implemented by distributing new payoff tables which are based on a smaller quantity of money. In particular, it is publicly announced at the end of period $t = T/2$ that all n firms receive new payoff tables. Again, each player is informed about his own new payoff table and those of the other $n-1$ players. We implement an *anticipated* monetary shock with this procedure because the firms get the new tables (with sufficient time for study) before they have to take their decisions in $T/2+1$, meaning that they know the other firms' new payoff tables and they know that others know this, and so on. We implement a *negative* monetary shock because the new payoff tables are based on a quantity of money M_1 which is smaller than the previous quantity $M_0 (M_1 = M_0/3)$. Finally, we implement an *exogenous* monetary shock because the firms get the new tables in $t = T/2$, irrespective of previous decisions. The parameters of the experiment imply a *unique* money-neutral equilibrium. Since the quantity of money falls by two-thirds, the price level should theoretically also fall by two-thirds. In particular, the average price taken over all n firms falls from 18 (in the pre-shock equilibrium) to 6 (in the post-shock equilibrium).

To investigate whether money illusion is a cause of nominal inertia, we chose the following design (see Figure 8.1). The first variation concerns the variation of the *representation of payoffs*. In the real representation, the payoff matrix shows *real* payoffs. That is, the numbers in the payoff matrix show how much the subjects will be paid at the end of a period for any feasible $p_i - P_{-i}$-combination. In the nominal representation, the payoff matrix shows nominal payoffs. Subjects must deflate nominal payoffs in order to determine what the corresponding real payoffs are, meaning they must divide the nominal payoff shown in the figure by the prevailing level of P_{-i}. Note that, with one exception, the payoff figures are completely identical in the two representations. All real payoffs are multiplied by the relevant average price P_{-i} in the nominal representation, while this is not the case in the real representation. We instructed the subjects who participated in the nominal treatment how to calculate real payoffs from nominal payoffs before the beginning

Representation of payoffs	Need to form expectations: Yes (Human opponents)	Need to form expectations: No (Computerized opponents)
Nominal	NH	NC
Real	RH	RC

Figure 8.1 Treatment conditions

of the experiment. Subjects had to solve several control questions to make sure that they knew how to perform these computations, and in addition they received a pocket calculator to facilitate their computations. All of the subjects solved all exercises successfully.

The second variation concerns *whether subjects need to form expectations* about the price choices of other firms (that is, whether there is strategic uncertainty). In the 'human opponents' treatments, subjects know that they interact with $n - 1$ other human subjects (see Figure 8.1). Subjects participating in these treatment conditions have to indicate their expectations EP_{-i} about the price level P_{-i} in each period. In the 'computerized opponents' treatments, subjects know that they play against $n - 1$ computers, and they know how these computers are programmed. In particular, the computers are programmed to simulate agents with perfect foresight. Thus, each subject i knows for certain that if i chooses price x, then the $n - 1$ computers are going to choose prices that result in a price level P_{-i} of y. This means that there is *no need* for subjects to form expectations in the computerized treatments. A subject's task is therefore reduced to an individual optimization problem. The experimental parameters are such that if subject i knows where the equilibrium is, he or she has no incentive whatsoever for not choosing the unique equilibrium.

Hypotheses

If one neglects disequilibrium play, as is routinely done in rational expectations models,[6] there should be no nominal inertia in all four cells of Figure 8.1. The reason is that these models assume money illusion and expectations formation to be irrelevant since the full rationality of all agents is assumed to be common knowledge. Therefore, nominal prices should adjust to the anticipated monetary shock instantaneously and equiproportionately. As a consequence, the anticipated monetary shock should be perfectly 'neutral' (that is, have no effect on the efficiency of the experimental economy) in all four treatments.

The deviation of post-shock nominal prices from equilibrium prices in the real representation with computerized opponents (RC) is a measure of individual-level irrationality which is unrelated to money illusion. For example, some subjects may be inattentive or confused by the monetary shock. In this case, nominal prices will not adjust to the nominal shock in cell RC instantaneously.

The effect of expectations formation is measured by the difference in adjustment speed between RH and RC. In both treatments, payoffs are represented in real terms. In RH, subjects have to form expectations about the effect of the monetary shock on other subjects' pricing decisions, whereas they do not have to do so in RC. If nominal prices exhibit more inertia in RH than in RC, it must be because some subjects expected that other subjects would not fully adjust nominal prices or because they are confused.

The effect of individual-level money illusion is measured by the difference in adjustment speed between NC and RC. In both cases, subjects do not have to form expectations about the decisions of other firms. The only difference between these two treatments is the nominal versus the real representation of payoffs. Therefore, the difference in adjustment speed between NC and RC measures how money illusion affects individual behaviour; that is, it measures the direct effect of money illusion.

The most interesting comparison occurs between NH and RH. Subjects have to form expectations about the pricing decisions of the other human subjects in both treatments. The only difference between these treatments is the nominal versus real representation of payoffs. In particular, the difference in adjustment speed between NH and RH measures the direct and indirect effects of money illusion.

Results

The main results are summarized in Figure 8.2, that shows average nominal prices in the four treatments. The data shown were generated by the decisions of 130 subjects who earned an average of $28. Each subject only participated in one of the four treatments. The first interesting point is that prices equilibrated quite nicely to the pre-shock equilibrium level of 18 in all four treatments.

In the real representation with computerized opponents (RC), *100 per cent of the subjects* (22 out of 22) *instantaneously and perfectly adjusted* prices to the shock. Therefore, the observed behaviour is perfectly in line with the

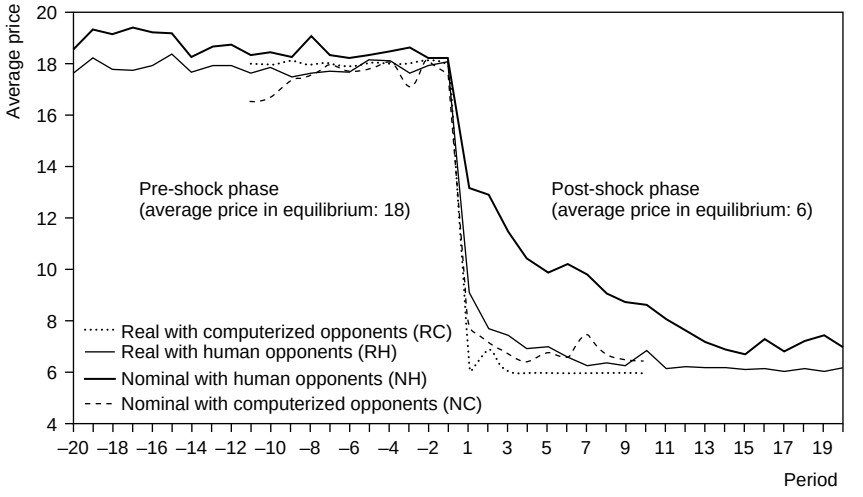


Figure 8.2 Average nominal prices

standard (macro-)economic theory prediction. As a consequence of this perfect price adjustment, the anticipated monetary shock is perfectly 'neutral' in this case. We conclude from this observation that the shock itself results in no noticeable confusion of subjects. However, it should be noted that the laboratory environment we implement is rather simple and easy to understand for subjects.

In the real representation with human opponents (RH), only 35 per cent of the subjects (14 out of 40) adjusted nominal prices to the shock instantaneously and perfectly, but most initial post-shock price choices were close to the equilibrium. The comparison of RC and RH shows that nominal prices exhibit significantly more inertia if subjects have to form expectations than when they do not (see Figure 8.2). A regression analysis shows that average prices in RH differ significantly from the equilibrium price level for two periods. We conclude that our human subjects are not able to solve the problem of coordinating expectations on the unique equilibrium perfectly. Put differently, the assumption of common knowledge of rationality does not seem to hold.

In the nominal representation with computerized opponents (NC), 79 per cent of the subjects (19 out of 24) instantaneously and perfectly adjusted prices to the shock. Taken together, the results from the treatments with computerized opponents indicate that there is only a small amount of money illusion at the individual level, but that there is no individual-level irrationality beyond that. Since price adjustment is only slightly slower in the NC than in the RC, we conclude that individual-level money illusion does not cause pronounced nominal inertia (compare the two dotted lines in Figure 8.2).

In the nominal representation with human opponents (NH), the adjustment of nominal prices to the monetary shock is *very sticky* (Figure 8.2). In particular, only 11 per cent of the subjects (5 out of 44) fully adjusted nominal prices to the equilibrium level in the first post-shock period and, as a consequence, the nominal price level fell by less than half of the predicted amount. According to our regression analysis, it takes 12 periods for full price adjustment in NH, whereas nominal prices equilibrate already after two periods in RH. The observed differences in adjustment speed in NH and RH also translate into different real effects of the monetary shock in the two treatments. For example, the average income loss is roughly twice as large in the NH as in the RH over the first 10 post-shock periods. Since the adjustment of nominal prices is much more sticky in the NH than in the RH, we conclude that the direct and indirect effects of money illusion are an important cause of nominal inertia (compare the two solid lines in Figure 8.2). As a consequence, the anticipated monetary shock is far from neutral in this environment.

A closer look at the expectations data reveals that the reason why nominal prices were much more sticky in the nominal representation (NH) than in the real representation (RH) is that expectations were much more sticky in NH than in RH. That is, subjects expected other subjects to choose high

prices in the nominal representation. Because subjects act in an individually rational manner (more than 80 per cent of subjects choose best replies to their expectations in NH and RH) and because of strategic complementarity, sticky expectations translate into sticky price choices.[7]

Asymmetric effects of positive and negative monetary shocks

So far, the results have shown that money illusion in fact causes nominal inertia and that the reason for this is that price expectations were much stickier in the nominal than in the real representation. But why was this so? Since we implement a *negative* monetary shock, the equilibrium price level must fall. By definition, high nominal payoffs prevail at high price levels. If subjects believed that high nominal payoffs 'look attractive' to other subjects, and if they believe that this causes other subjects to choose high prices in the post-shock phase, they respond rationally by also choosing high prices in the post-shock phase. To test for this hypothesis, we implemented a *positive* monetary shock with human opponents (NH, RH). If our hypothesis about the cause of the stickiness of expectations is correct, prices should adjust much more quickly after the positive than after the negative shock, because equilibrium price levels have to rise with a positive shock and therefore subjects have to adjust their price choices in the direction of high 'attractive' nominal payoffs. The experiments were run with an additional 96 subjects and strongly confirm this hypothesis (see Fehr and Tyran, 2001 for details). We observe a pronounced *asymmetry* in nominal inertia; that is, a much quicker convergence to the equilibrium after a positive shock than after a negative shock in the NH. This finding also suggests that money illusion may provide a micro-foundation for the asymmetrical real economic effects of positive and negative monetary shocks which seem to have been observed (for example, Cover, 1992; Peltzman, 2000).

Strategic complementarity as a cause of nominal inertia

The explanation provided above on why money illusion causes nominal inertia is based on the idea that money illusion systematically affects price expectations, and decision-makers react rationally to these expectations. Sticky expectations after the negative shock in NH translated into sticky price choices because of strategic complementarity. According to this reasoning, strategic complementarity plays a key role in determining whether sticky expectations translate into aggregate-level effects of money illusion. We implemented a negative shock with a nominal representation (NH) and either strategic complements (a positive slope of the reaction function) or strategic substitutes (a negative slope of the reaction function) to test for the role of strategic properties. Rational agents expecting under-adjustment by illusion-prone agents tend to imitate the behaviour of illusion-prone agents if strategic complements prevail, but to *compensate* their behaviour if strategic substitutes prevail. Therefore, subjects prone to money illusion should have

a disproportionately large effect on the aggregate price level if strategic complements prevail but a disproportionately small effect if strategic substitutes prevail. Our results (from an additional 76 subjects) support this hypothesis.

Figure 8.3 shows average prices and average price expectations in the two treatments. As can be seen, prices and expectations converge nicely to the equilibrium in the pre-shock phase (periods 1–15). In response to the anticipated negative monetary shock at the beginning of period 16, expectations differ dramatically across treatments. While expectations are very sticky with strategic complements (see upper dotted line), average expectations remain very close to the predicted equilibrium with strategic substitutes. While expectations only converge slowly to the equilibrium with strategic complements, they remain in equilibrium with strategic substitutes from the second post-shock period on. Fehr and Tyran (2002) provide a detailed account of the observation that strategic properties have such a marked effect on expectations. Simulation results suggest that the expectations formation process is quite different in the two conditions. In particular, expectations with strategic substitutes are much more in line with the predictions of 'rational expectations'. Because expectations are almost instantaneously in equilibrium (and because agents choose best replies to their expectations), there is no nominal inertia and the anticipated monetary shock is almost neutral if subjects' actions are strategic substitutes (see Fehr and Tyran, 2002 for the details).

It is important to note that all experimental studies discussed above had unique equilibria. We found that money illusion systematically affects

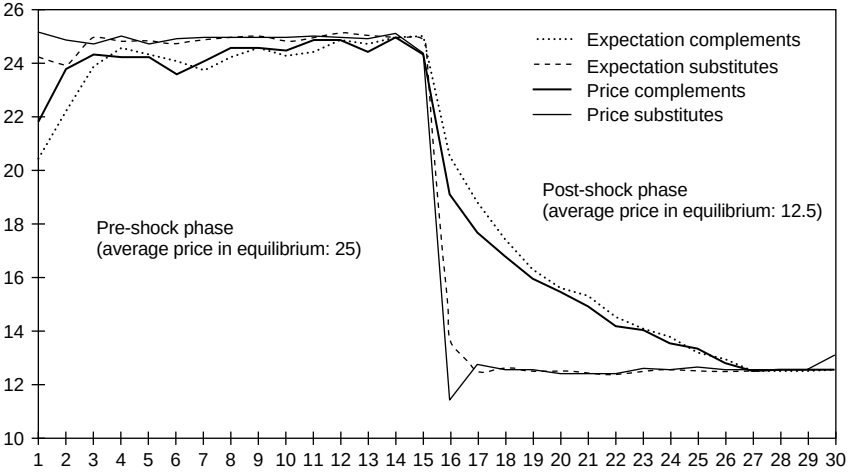


Figure 8.3 Average expectations and prices with strategic complements and substitutes

convergence to this unique equilibrium in such an environment because it shapes expectations. Eventually, however, prices converged to the unique equilibrium in all cases. As a consequence, the experiments discussed so far demonstrate a short-run real effect of money illusion which is more or less pronounced, but they do not provide evidence for permanent effects of money illusion.

5 Money illusion and coordination failure

We now analyse whether expectations can induce permanent effects of money illusion in an environment with strategic complements. To do so, we chose an experimental design that is very similar to but simpler than that discussed in section 4. The main difference to the design described there is that there is no monetary shock but there are *multiple* equilibria. Multiple equilibria prevail if the degree of strategic complementarity is (locally) extreme (see Cooper, 1999); such a situation is illustrated in Figure 8.4. The equilibria we implement are Pareto-ranked. Equilibrium A (at a low price level) yields a higher real payoff for all players than equilibrium C, the lowest real payoff. As in section 4, we capture the impact of money illusion by comparing behaviour in a treatment condition in which payoff information is provided in real terms with the behaviour in a treatment condition in which payoff information is provided in nominal terms. In the nominal representation, the *nominal* payoffs in the efficient equilibrium A are lower than those in the inefficient equilibrium C. Thus, if subjects take nominal payoffs as a proxy for real payoffs they may mistakenly prefer being in the inefficient equilibrium C and choose high prices accordingly. Our results clearly illustrate that money illusion can be a powerful source of coordination failure causing permanent real effects.

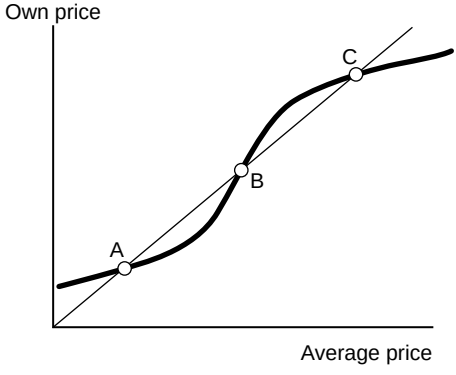


Figure 8.4 Best reply structure with multiple equilibria

Experimental design and hypotheses

As in the design described in the previous section, subjects had to simultaneously choose a price $p_i \in \{1, 2, \ldots, 30\}$. Each subject's real payoff depended only on his own price p_i and on the average price P_{-i} of the other $n-1$ players, and the payoffs were represented in a payoff matrix (see Fehr and Tyran, 2004, for details). The three equilibria of the game are described in Table 8.1. To play equilibrium A, each subject had to choose a price of $p_i = 4$. Because the game is symmetric, this led to a price level of $P_A = 4$, and resulted a real payoff of $\pi_A = 28$ for each player. In equilibrium B the price was $P_B = 10$ causing a real payoff of 5. In equilibrium C, the price level was $P_C = 27$ and each player earned a real payoff of 21. Each subject i had a unique and weakly increasing best reply for every given level of P_{-i}. Since the game is symmetric, this game's equilibria are located at the intersection of the best reply function with the 45-degree line in the (p_i, P_{-i})-space (see Figword 8.4). If subjects have adaptive expectations and play a best reply to their expectation, equilibria A and C are stable, and equilibrium B is unstable.

The main purpose of our design was to create a conflict between two potentially important equilibrium selection principles – the principle of real payoff dominance and the principle of nominal payoff dominance. To examine the importance of nominal payoff dominance, we chose the parameters in such a way that the real payoff is highest for each player in equilibrium A but the nominal payoff is highest in equilibrium C (see Table 8.1). Thus, the principle of real payoff dominance predicts that equilibrium A will be selected while the principle of nominal payoff dominance predicts that equilibrium C will be selected. If we indeed observe that players permanently coordinate on equilibrium C we have not only evidence that nominal payoff dominance is stronger but we also have evidence for the striking claim that money illusion may have permanent real effects.

Our price-setting game was implemented in four different treatment conditions analogous to the lines explained in section 4 (see Figure 8.1). The treatments differed with regard to the nominal versus real presentation of the payoffs and whether subjects had to form expectations (because they played against other subjects) or whether they played against $n-1$ preprogrammed computers. In the latter case, subjects did not have to form

Table 8.1 The equilibria in the price-setting game

Equilibrium	*Equilibrium price level*	*Real equilibrium payoff*	*Nominal equilibrium payoff*
A	$P_A = 4$	$\pi_A = 28$	$P_A\pi_A = 112$
B	$P_B = 10$	$\pi_B = 5$	$P_B\pi_B = 50$
C	$P_C = 27$	$\pi_C = 21$	$P_C\pi_C = 567$

expectations because they knew the computers' response to each of their feasible price choices. To maximize their payoffs, subjects had to solve an individual optimization problem taking the computers' aggregate response into account. Therefore, the treatments with computerized opponents measure the extent to which subjects are able to solve this optimization problem by choosing the efficient equilibrium A.

According to Table 8.1, it is obvious that equilibrium A dominates the other two equilibria in real terms. However, subjects may not be able to play the best equilibrium immediately. They may have to learn to play the best equilibrium when facing human opponents or their optimal strategy when facing computerized opponents. We repeated the same game for $T = 30$ periods in each treatment condition for this reason. When subjects faced human opponents, the group composition remained constant throughout the 30 periods. Subjects were informed in all conditions about the actual average price of the other players, P_{-i}, at the end of each period and about their real payoff.

The overall purpose of our treatment conditions was to isolate the role of money illusion as an equilibrium selection device from other boundedly rational forms of equilibrium selection. The two major conditions in our design are the Real treatment with human opponents (RH) and the Nominal treatment with human opponents (NH). The difference between these two conditions informs us about the overall effect of money illusion on equilibrium selection.

Our particular interest concerns the role of *expectations* in the selection effects of money illusion. In principle, money illusion can affect equilibrium selection in two ways. First, there may be direct effects on equilibrium selection: subjects may play the inefficient equilibrium C because they are prone to individual-level money illusion. Second, indirect effects may arise from expectations about other players' money illusion. Even if no player exhibits individual-level money illusion, most subjects may nevertheless have an incentive to play the inefficient equilibrium C if they expect that a sufficient number of other players suffer from money illusion and that these will, therefore, play equilibrium C. Our treatments with computerized opponents enable us to isolate the extent to which individual-level money illusion directly affects equilibrium selection.

Results

In total, 174 subjects participated in our experiments. Subjects were randomly allocated to groups of $n = 5$ or $n = 6$ players. They received written instructions explaining the experimental procedures and nominal or real payoff matrices depending on the treatment condition. The calculation of real payoffs from nominal payoffs shown on the payoff matrix was carefully explained in the NC and the NH. Subjects had to choose a

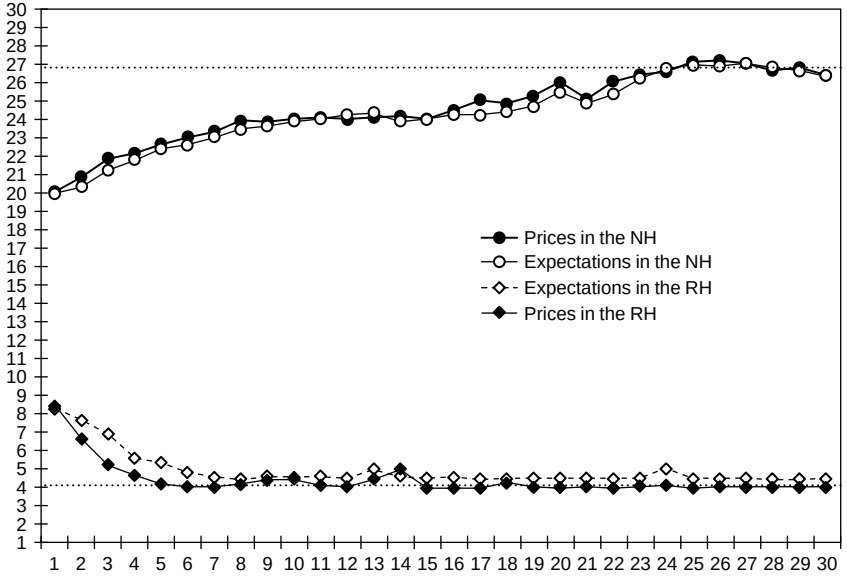


Figure 8.5 Average prices and expectations in the treatments with human opponents

price $P_i \in \{1, 2, \ldots, 30\}$ in each period. In addition, they had to indicate their expectation of EP_{-i} in each period in the treatments with human opponents.

Our first main result concerns the comparison of the NH and the RH. The vast majority of the subjects converge to the *inefficient* equilibrium C in the Nominal treatment with Human opponents (NH), whereas almost all subjects quickly converge to the *efficient* equilibrium A in the Real treatment with Human opponents (RH). Figure 8.5 shows that there is already a large gap in average prices across treatments in period 1 – in the NH the average price is 20.1 in the first period whereas in the RH it is 8.4. The Mann–Whitney test yields a highly significant difference ($p < .0001$). Moreover, the average price quickly converges towards the efficient equilibrium $P_A = 4$ in the RH whereas a slow but steady convergence to the inefficient equilibrium $P_C = 27$ occurs in the NH.

Individual price choices in the NH and the RH differ radically. Not a single subject chose the efficient equilibrium in the NH throughout the 30 periods, while 64 per cent of the subjects already chose $P_A = 4$ in period 1 in the RH. More than 90 per cent of the subjects always chose the efficient equilibrium in the RH after period 7. In contrast, there is much disequilibrium play in the NH and a relatively slow convergence to the inefficient equilibrium $P_C = 27$ occurs. Eighteen per cent of the subjects chose $P_C = 27$ in period 1 and they gradually increased to 84 per cent in period 30.

This divergence between the NH and the RH is reflected in the real payoffs the subjects earned. They earned considerably less in the NH than in the RH in all periods. Recall from Table 8.1 that the real equilibrium payoff in the efficient equilibrium A is 28 while in the inefficient equilibrium C it is only 21. There was rarely a period in which subjects did not earn an average of at least 10 units less in the NH. This indicates that the miscoordination in the NH goes beyond the fact that subjects coordinated on an inefficient equilibrium. The large payoff difference is partly caused by the larger incidence of disequilibrium play in the NH.[8]

The striking price divergence across the NH and the RH suggests that money illusion has powerful effects on equilibrium selection. The mere fact of payoffs being represented in nominal terms induces subjects to predominantly choose the equilibrium with the higher nominal but the lower real payoff. This fact is to be explained by the movement of price expectations across treatments. The average price path closely parallels subjects' average expectations EP_{-i} (see Figure 8.5) in both the RH and in the NH. Since subjects almost always played a best reply to their expectation EP_{-i}, this expectation is a decisive determinant of subjects' price choices. It is therefore interesting to know that subjects' price expectations already differed strongly in period 1: EP_{-i} was 20.0 in the NH whereas they expected a value of 8.2 in the RH. In the NH not a single subject (out of 77) expected an equilibrium of $EP_{-i} = 4$ in period 1. In contrast, 48.1 per cent (25 out of 52) held equilibrium expectations of $EP_{-i} = 4$ in the RH.

So far, our analysis suggests that the nominal representation of payoffs causes significantly higher price expectations, which in turn induces subjects to choose significantly higher prices in the NH. This raises the question whether there were indeed subjects who failed to see through the veil of money or whether the expectations of higher prices in the NH were solely rooted in subjects' beliefs about other players' money illusion. We obtain the following result for the existence of individual-level illusion: only a minority of the subjects initially play the efficient equilibrium in the Nominal treatment with computerized opponents (NC) whereas a large majority of subjects play the efficient equilibrium from the beginning in the Real treatment with computerized opponents (RC). However, the differences between the NC and the RC decrease and lose significance over time. Thus, the evidence suggests that the nominal representation causes significantly more problems for the subjects in solving the individual optimization problem. This provides direct evidence for individual-level money illusion. Yet, over time, subjects learn to pierce the veil of money better and to solve the optimization problem in the NC roughly in the same way as in the RC.

We examine next how strategic interaction in the RH affects individual irrationality other than money illusion. We find that strategic interaction increases the frequency with which the *efficient* equilibrium is played in the

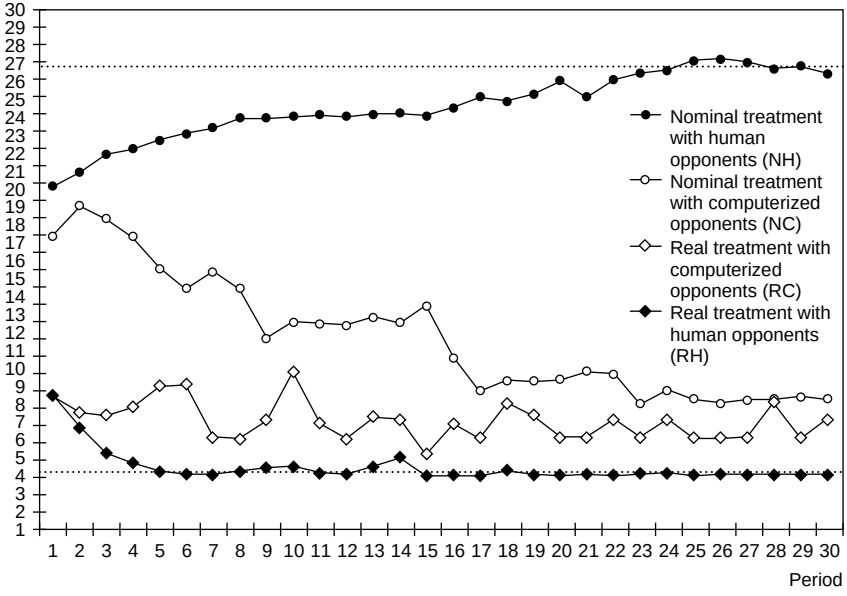


Figure 8.6 Average prices across all treatments

treatments with a real payoff representation and, eventually, removes almost all inefficiencies.

Figure 8.6 shows that the average price in the RC and the RH is almost identical in period 1. Following this period, the average price quickly converges to the efficient equilibrium in the RH while it fluctuates between 2 and 4 units above the efficient equilibrium in the RC. Thus, even though the difference between the RC and the RH is small, it persists over time. This indicates that there is a small amount of individual-level irrationality in the RC, which is largely absent in the RH. Interestingly, when payoffs are represented in real terms, strategic interactions do not magnify but remove the impact of individual-level bounded rationality on miscoordination.

In contrast, we find that strategic interaction causes a large increase in the frequency with which the *inefficient* equilibrium C is played in the treatments with a nominal payoff representation, and it completely removes the play of the efficient equilibrium from the beginning. Figure 8.6 indicates that the average prices in the NC and the NH are relatively close together in the first two to three periods. However, whereas the average price rises steadily in the NH, it falls in the NC. The reason for the diverging price movements is that subjects learn to choose the efficient equilibrium in the NC whereas groups increasingly coordinate on the inefficient equilibrium in the NH.

In our view, the comparison between the NC and the NH is exciting because it suggests that most subjects do learn to play the efficient equilibrium when they are provided with individual learning opportunities and when they are not trapped by the powerful attraction of an inefficient equilibrium. Thus, individual learning largely removes the veil of money that blinds subjects in a non-strategic environment. However, when subjects have to form expectations because they play the strategic game with other humans, the level of money illusion which initially exists attracts subjects towards the inefficient equilibrium from which no escape seems possible.

6 Summary and conclusion

In principle, money illusion could explain nominal inertia and, as a consequence, explain the non-neutrality of money. However, mainstream economists have dismissed this 'psychological' explanation for two reasons. First, money illusion is rejected *a priori*, simply because it contradicts the basic rationality assumptions of economics. Second, there was no convincing evidence for the existence and relevance of money illusion. Recently, however, Shafir, Diamond and Tversky (1997) have provided evidence from questionnaire surveys suggesting that money illusion is an important phenomenon at the individual level. However, some economists tend to argue that the evidence that such surveys yield about the existence of money illusion is weak and irrelevant. It is weak because people have no real incentive to think about their decisions, and it is irrelevant because individual-level effects may disappear with interaction. Experimental methods allow a response to these criticisms, as the actual behaviour of real people is observed under controlled conditions in experimental studies, and these people are motivated by economic incentives.

The claim that money illusion must be irrelevant because individual-level effects are small and transitory is seriously misleading in our opinion, as money illusion can have important indirect effects by shaping expectations. Forming expectations is necessary for making optimal decisions in a strategic environment. Strategic complementarity prevails if agents have an incentive 'to follow the crowd', and strategic complementarity may multiply the effects of expectations shaped by money illusion.

We developed an experimental design to investigate whether money illusion is a cause of nominal inertia. Our results show that money illusion can have massive aggregate-level effects under conditions of strategic complementarity, even if individual-level money illusion is small. Our results suggest that money illusion induces expectations of sticky price adjustment following a negative monetary shock and that these 'sticky' expectations translate into sticky pricing decisions if strategic complementarity prevails. Furthermore, the results indicate that money illusion induces the asymmetric effects of positive and negative nominal shocks. In particular, price

adjustment is much faster and real effects are much less pronounced after a positive than after a negative nominal shock. Finally, it is shown that strategic complementarity is a key element in understanding the causes of nominal inertia. In particular, we show that nominal inertia is much more pronounced if strategic complementarity prevails than if strategic substitutability prevails.

We investigate potential long-run effects of money illusion in an environment with multiple Pareto-ranked equilibria in a related set of experiments. Multiple equilibria arise if the degree of strategic complementarity is (locally) extreme. Our results show that money illusion can coordinate expectations on an inefficient equilibrium in which agents get locked in.

We believe that these results constitute important insights into the aggregate-level effects of money illusion and that the attainment of these insights seems almost impossible without controlled laboratory experiments. The experimental method makes it possible to precisely identify the conditions under which rational expectations models are correct, and the conditions under which they fail to capture important economic forces and facts. Since we show that strategic complementarity is a key determinant of the aggregate-level effects of money illusion, and since strategic complementarity seems to be an important feature of reality (Cooper and Haltiwanger, 1996), our experiments suggest that it is time to take money illusion seriously in macroeconomics.

Notes

1 There was a third group of respondents that was asked whether Ann or Barbara is doing better in economic terms; 71 per cent answered that Ann is in fact doing better in economic terms.

2 Before the advent of the rational expectations revolution, the assumption of money illusion was usually not justified explicitly. Money illusion was often casually invoked, probably because it seemed such a natural assumption at the time. For example, Milton Friedman (1968) proposed a theory of monetary non-neutrality that is based on a misperception of real wages by workers. According to Friedman, when the money supply rises unexpectedly, the price level rises, pushing down the real wage. Employers hire more because the cost of labour has fallen. Employees are willing to work more because they focus on the nominal wage and infer (incorrectly) that the reward for work has risen. While Friedman did not explain this asymmetry between workers and firms, it seemed natural at the time to assume that workers at least partly ignore the effect of a price level increase on real wages. This ignorance was the centrepiece of Friedman's proposed explanation for the short-run Philips curve.

3 The rationality requirements are much higher for the attainment of an equilibrium than they are for individual maximization as pointed out by Arrow (1987), for example. Attainment of the equilibrium involves 'an informational burden of an entirely different magnitude than simply optimizing at known prices' (1987: 201).

4 Motolese (2003) provides an alternative approach in which endogenously heterogeneous beliefs can cause monetary non-neutrality.

5 See Conlisk (1996) for a more general discussion of the relation between limited rationality and limited information.
6 According to Muth (1961: 316), rational expectations are equilibrium expectations: 'I should like to suggest that expectations, since they are informed predictions of future events, are essentially the same as the predictions of the relevant theory.'
7 Note that we asked each subject *i* to indicate first-order expectations EP_{-i} about the average behaviour of the other $-i$ firms. To keep the experiment simple, we did not ask subjects to indicate second-order expectations [i.e., *i*'s expectations about *j*'s expectation about what everybody else is going to do $E_i(E_jP_{-j})$], or expectations of an even higher order. Data on higher-order expectations would be necessary to analyze in greater detail why first-order expectations are biased by money illusion. Suppose, for example, that *i*'s first-order expectations are biased but *i*'s second-order expectations are not. In this case, *i* expects the other $-i$ players to be prone to money illusion (or to be otherwise irrational), but does not believe that these players expect others (or expect others to expect again others, and so on) to be prone to money illusion. In contrast, suppose *i*'s first-order and second-order expectations are biased. Then, *i* expects *j*'s expectations to be partly shaped by $-j$'s money illusion (or $-j$'s expectation about others money illusion, and so on). While a more pronounced bias in first-order expectations in the NH compared to the RH does not allow us to determine exactly at which level money illusion distorts behaviour, it clearly indicates that money illusion shapes expectations at some level.
8 There were several cases where subjects initiated disequilibrium by deliberately trying to push the group towards the efficient equilibrium with the choice of low prices.

References

Adam, K. (2003) *Optimal Monetary Policy with Imperfect Common Knowledge*, Working Paper no. 2003/12, Center for Financial Studies, University of Frankfurt, April.

Akerlof, G.A., W.T. Dickens and G.L. Perry (1996) 'The Macroeconomics of Low Inflation', *Brookings Papers on Economic Activity*, vol. 1, pp. 1–76.

Arrow, K.J. (1987) 'Rationality of Self and Others in an Economic System', in R.M. Hogarth and M.W. Reder (eds), *Rational Choice: The Contrast between Economics and Psychology* (Chicago: University of Chicago Press), pp. 201–15.

Ball, L., N.G. Mankiw and R. Reis (2003) *Monetary Policy for Inattentive Economies*. NBER Working Paper no. 9491.

Belongia, M.T. (1996) 'Measurement Matters: Recent Results from Monetary Economics Reexamined', *Journal of Political Economy*, vol. 104(5), pp. 1065–83.

Blinder, A.S., E.D. Canetti, D.E. Lebow and J.B. Rudd (1998) *Asking About Prices. A New Approach to Understanding Price Stickiness* (New York: Russell Sage Foundation).

Camerer, C.F. (2003) *Behavioral Game Theory* (Princeton, NJ: Princeton University Press).

Conlisk, J. (1996) 'Why Bounded Rationality?', *Journal of Economic Literature*, vol. 34(2), pp. 669–700.

Cooper, R.W. (1999) *Coordination Games: Complementarities and Macroeconomics* (Cambridge: Cambridge University Press).

Cooper, R.W. and J. Haltiwanger (1996) 'Evidence on Macroeconomic Complementarities', *Review of Economics and Statistics*, vol. 78(1), pp. 78–93.

Cooper, R.W. and A. John (1988) 'Coordinating Coordination Failures in Keynesian Models', *Quarterly Journal of Economics*, vol. 103(3), pp. 441–63.

Cover, J.P. (1992) 'Asymmetric Effects of Positive and Negative Money-Supply Shocks', *Quarterly Journal of Economics*, vol. 107(4), pp. 1261–82.
Duffy, J. (1998) 'Monetary Theory in the Laboratory', *Federal Reserve Bank of St. Louis Review*, Sept./Oct., pp. 9–26.
Fehr, E. and J.-R. Tyran (2001) 'Does Money Illusion Matter?', *American Economic Review*, vol. 91(5), pp. 1239–62.
Fehr, E. and J.-R. Tyran (2002) *Limited Rationality and Strategic Interaction. The Impact of the Strategic Environment on Nominal Inertia*, Working Paper no. 2002–25, Department of Economics, University of St Gallen, November.
Fehr, E. and J.-R. Tyran (2004) *Money Illusion and Coordination Failure*, CESifo Working Paper no. 1141, University of St Gallen, February.
Fischer, S. (1977) 'Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule', *Journal of Political Economy*, vol. 85(1), pp. 191–205.
Fisher, I. (1928) *The Money Illusion* (Toronto: Longmans).
Friedman, M. (1968) 'The Role of Monetary Policy', *American Economic Review*, vol. 58(1), pp. 1–17.
Haltiwanger, J.C. and M. Waldman (1985) 'Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity', *American Economic Review*, vol. 75(3), pp. 326–40.
Haltiwanger, J.C. and M. Waldman (1989) 'Rational Expectations and Strategic Complements: The Implications for Macroeconomics', *Quarterly Journal of Economics*, vol. 104(3), pp. 463–84.
Howitt, P. (1989) 'Money Illusion', in J. Eatwell, M. Milgate and P. Newman (eds), *Money* (New York and London: W.W. Norton), pp. 244–7.
Hume, D. (1752) 'Of Money; Of Interest', reprinted in E. Rotwein (ed.) (1970), *Writings on Economics* (Madison: University of Wisconsin Press).
Kahn, S. (1997) 'Evidence of Nominal Wage Stickiness', *American Economic Review*, vol. 88(5), pp. 993–1008.
Lucas, R.E. Jr (1972) 'Expectations and the Neutrality of Money', *Journal of Economic Theory*, vol. 4(2), pp. 103–24.
Mankiw, N.G. (1985) 'Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly', *Quarterly Journal of Economics*, vol. 100(2), pp. 529–37.
Mankiw, N.G. and R. Reis (2002) 'Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve', *Quarterly Journal of Economics*, vol. 117(4), pp. 1295–328.
Motolese, M. (2003) 'Endogenous Uncertainty and the Non-Neutrality of Money', *Economic Theory*, vol. 21, pp. 317–45.
Muth, J.F. (1961) 'Rational Expectations and the Theory of Price Movements', *Econometrica*, vol. 29(3), pp. 315–35.
Peltzman, S. (2000) 'Prices Rise Faster than They Fall', *Journal of Political Economy*, vol. 108(3), pp. 466–502.
Shafir, E., P.A. Diamond and A. Tversky (1997) 'On Money Illusion', *Quarterly Journal of Economics*, vol. 112(2), pp. 341–74.
Sims, C.A. (2003) 'Implications of Rational Inattention', *Journal of Monetary Economics*, vol. 50, pp. 665–90.
Taylor, J.B. (1979) 'Staggered Wage Setting in a Macro Model', *American Economic Review*, vol. 69(2), pp. 108–13.
Tobin, J. (1972) 'Inflation and Unemployment', *American Economic Review*, vol. 62(1), pp. 1–18.

Tversky, A. and D. Kahneman (1981) 'The Framing of Decisions and the Psychology of Choice', *Science*, vol. 211, pp. 453–8.

Tversky, A. and D. Kahneman (1986) 'Rational Choice and the Framing of Decisions', *Journal of Business*, vol. 49(4), pp. 251–78.

Woodford, M. (2003) 'Imperfect Common Knowledge and the Effects of Monetary Policy', in P. Aghion, R. Frydman, J. Stglitz and M. Woodford (eds), *Knowledge, Information and Expectations in Modern Macroeconomics: Essays in the Honor of Edmund S. Phelps* (Princeton, NJ: Princeton University Press).

9

Utility-Based Altruism: Evidence from Experiments*

Alexander Kritikos and Friedel Bolle
European University Viadrina, Frankfurt (Oder), Germany

1 Introduction

This chapter seeks to provide new evidence on economic approaches to the issue of altruism based on Dictator Game experiments, focusing on experiments where the information status of the recipient is variable.

Altruistic behaviour has received attention in both economics and psychology.[1] Also called 'pro-social behaviour' in psychological terminology, altruistic behaviour represents any behaviour which has to do with sharing a pie (as in the Dictator Game), or helping other people. Altruistic behaviour can be observed in many circumstances, ranging from everyday situations to situations of extreme distress.

Typically, psychological approaches to altruism tend to focus more on the processes leading to altruistic or non-altruistic behaviour while economic approaches tend to focus more on the interpretation of human decisions as being motivated either by altruism or otherwise. These differences in approach are apparent when we compare the design and aims of economic with psychological experiments on altruistic behaviour.

In economics the most prominent experiment used to test the existence of altruism is the Dictator Game. In this game person A, the dictator, can decide how to share a pie between himself and person B. Since the recipient cannot react explicitly to the dictator's decision, the dictator can only be influenced implicitly (if at all) by the recipient. Thus, the anonymously played Dictator Game can be used to test altruistic motives, since person A can

* This paper was presented at the Congress of the International Economic Association 2002 in Lisbon. We would like to thank Simon Gaechter, Werner Gueth, Gunda Laasch-Wrobel, Axel Ockenfels and in particular the two editors of this volume for their helpful comments. We further thank the German Science Foundation (Grant Bo 747/5-1) for financial support.

express his willingness to reduce his own level of consumption in favour of person B.

A core result of Dictator Game experiments (see for example, Forsythe *et al.*, 1994) is that approximately one-third of the participants give nothing, while roughly two-thirds give somewhere between 20 and 50 per cent of a pie of $10. This finding was supported by later experiments by several scholars, including Hoffman *et al.* (1994), Camerer and Thaler (1995), Bolton and Zwick (1995), Eckel and Grossman (1996), Andreoni and Miller (2002) and Andreoni and Vesterlund (2001), where the modal and average offer to the recipient varied between 20 and 25 per cent.

In contrast, classical psychological experiments on altruistic behaviour focus on the conditions under which a certain person is willing to help another in need. They aim at finding out to what extent a person's willingness to help depends on conditions such as being the only person around, or being one of two or more persons, or being with friends or strangers, and so on (Latané and Rodin, 1969; Wilson, 1976; Hansson and Slade, 1977). They reveal that altruistic behaviour depends strongly on these differing conditions and contexts.

Today the existence and relevance of altruism is accepted in both professions. However, the impact of altruistic behaviour has rarely been considered in more general models,[2] possibly because altruism is an elusive concept not only in theory, as Simon (1993) has highlighted, but also in attempts to isolate altruistic behaviour in controlled experiments. From a theoretical viewpoint, it is difficult to define the utility of the recipient and the utility and cost of altruistic choices for the donator; which in turn makes it difficult to judge the donator's motivation in making such a choice. Despite these difficulties however, some of the attempts at modelling have proved promising and insightful.

With some exceptions, economic approaches to altruistic choices have focused on material outcomes for donators and recipients. According to these models the donator aims at increasing the income of the recipient (see for instance, Collard, 1978; Bester and Gueth, 1998; Andreoni and Miller, 2002). It is mainly in Becker's (1974) model and in models concerning intergenerational transfers, starting with Barro (1974) where it is assumed that altruistic moves aim at increasing the utility of the recipient. Clearly, some economists fairly early on focused on psychological dimensions in their models.

In this chapter we further investigate these alternative forms of modelling altruistic behaviour. In order to better understand whether it is the utility or the income of the recipient which the donator aims to increase by his altruistic choice, we have developed a modified Dictator Game where the recipient has incomplete information about the size of the pie. We will use the Standard Dictator Game (as spelt out by Forsythe *et al.*, 1994) as a benchmark and compare behaviour in the usual Dictator Game with the modified game.

In section 2 we very briefly describe the income-based and the utility-based approaches to altruism and discuss the implications of the Dictator Game under incomplete information for both approaches. In section 3 we describe the setting of two Dictator Game experiments, one under complete information and the other under incomplete information. In the latter experiment, the dictator knows the size of the pie while the recipient knows only the probability distribution of the two potential pie sizes. In section 4 we derive some predictions and present the results in section 5. Those dictators who received large pies either offered the recipient close to an equal split of the pie (and, thus, revealed the true size of their pie) or pretended to have received a small pie and offered the recipient half or less of the small pie. This behaviour can be explained only by altruism as deduced from the utility-based approach. Section 6 provides summary comments.

2 A dictator experiment with an uninformed recipient

As noted above, there are essentially two main approaches to altruism in economic theory: the income-based approach and the utility-based approach. In most analyses of altruism, it is the former approach which is followed – it is assumed that a person's utility is influenced either by the other persons' consumption of goods or by the other persons' income:

$$U_i = V_i(x_1, \ldots, x_i, \ldots, x_n), \quad i = 1, \ldots, n,$$
$$\text{with } x_j = \text{income (or consumption) of person } j \tag{1}$$

where x_i represents i's consumption, and $x_1, \ldots, x_n$ represent the consumption of individuals j with whom altruist i interacts. Given (1) and $\partial V_i/\partial x_j > 0$, i's utility is increased if j enjoys a higher income. If $\partial V_i/\partial x_j < 0$ then i wants j to have a lower income.

Gary Becker (1974) has proposed a utility-based setting:

$$U_i = U_i(x_i, U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_n), \quad i = 1, \ldots, n \tag{2}$$

where $U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_n$ represent the utilities of individuals j with whom altruist i interacts. If $\partial U_i/\partial U_j > 0$ then i's utility is increased if j's utility ($j \neq i$) is increased or i is 'happy' if j is 'happy'. If $\partial U_i/\partial U_j < 0$ then i dislikes j. Note that the function $U_i(.)$ may implicitly contain additional parameters describing the environment and the specific situation.

The main reason why economists prefer the income-based approach, however, is the tractability of the model (see also Bolle, 1991). If applied to the Standard Dictator Game under complete information, both approaches lead to the same predictions about behaviour. Matters change in Dictator Games where the information status of the recipient is variable. In the usual Dictator

Game under complete information, person 1 (the dictator) is endowed with a known amount of money p which he can divide arbitrarily between himself by keeping x_1 and person 2 (the recipient) by transferring x_2. In the experiment we will present here, the Dictator Game varies insofar as the dictator is endowed with an amount p which is not known to the recipient. She only knows that $p = p_S$ with probability α and $p = p_L$, with $1 - \alpha$, and she knows $p_S < p_L$.[3] Further below we will concentrate on those dictators who received a large pie.

Under the income-based approach to altruism, it is clear that the dictator will offer the recipient the same amount x_2 and keep $x_1 = p - x_2$ for himself, irrespective of the recipient's information status about the pie size p.

If the utility-based approach to altruism is applied, the dictator will care whether the recipient's belief about p affects her utility. As a first step, under incomplete information, it will be necessary to explain the formation of consistent expectations and then to solve the resulting systems of equations. This allows us to derive predictions about the dictator's expected behaviour under incomplete information (for a detailed analysis see, Bolle and Kritikos, 2004. In this chapter we provide only a non-formal description).

We start with the expectations of the recipient. Standard Dictator Game experiments show that the recipient has expectations about the transfer x_2 that she would receive under complete information. It is reasonable to assume that she expects no dictator to give more than half of his endowment. By giving more than $p_S/2$ the dictator would reveal that he has the larger pie ($p = p_L$). By proposing less than $p_S/2$ he can make the recipient believe that $p = p_S$, at least with a certain degree of probability.

Earlier psychological research on equity theory (see for example, Walster *et al.*, 1978) and on the theory of justice (for example, Mikula, 1983; Reis, 1983) may now allow for the assumption that there are two dimensions which might increase the recipient's utility, the absolute and the relative amount the dictator is transferring to the recipient. With respect to the relative amount x_2/p it is further necessary to define what amount will increase the utility of the recipient. Among the existing criteria for distributive justice (Mikula, 1983; Konow, 2001), the norm of equality seems to be the most appropriate with respect to the dictator game. If we apply this to the Dictator Game under incomplete information it would mean that the recipient's utility increases the closer the amount appears to be to an equal split of the pie[4] and the higher the amount (in absolute terms) transferred to her.

Given that the distribution of p is common knowledge and given that the expectations of the recipient are consistent with those of the dictator, we expect to observe four different types of dictators, if endowed with a large pie p_L. The first type of dictator would be expected to transfer more than $p_S/2$, indicating that he is endowed with p_L. Since he is ready to reveal his type, he will give the same amount as under complete information. Of course,

the borderline between this first type and the other types is not exogenously given but determined by an analysis of the situation.

The second type of dictator (who received a large pie) will transfer $x_2 \leq p_S/2$ under incomplete information, although he would have proposed more than $p_S/2$ under complete information. He will increase his utility by reducing x_2 below $p_S/2$ because he expects that the recipient's utility can be increased by making her believe that she received a substantial share of the small pie. Hence, the utility of the recipient and, consequently, the utility of the dictator will increase.

For the second type, x_1 and x_2 are not equal to their optimal values under complete information. Those dictators (who would have given more than half of the small pie in the complete information setting) have to compare the 'indirect' increase in the recipient's utility from reducing their transfer to less than half the small pie (but close to the seemingly equal split of the small pie), with the 'direct' decrease in their utility arising from reducing their transfer below the optimal level. It is reasonable to expect that the decrease in 'direct' utility is greater the more x_2 has to be reduced under incomplete information in comparison to the optimal x_2 under complete information. So it makes sense to assume that dictators who are ready to reveal the true size of their pie are those types who give most under complete information.

However, it is still unclear how much dictators pretending to have received the small pie are expected to propose. Under complete information, it would be optimal to offer $x_1 > p_S/2$, under incomplete information the dictator would like to reduce his proposal as little as possible. Confronted with a naïve recipient, he would prefer to transfer an amount exactly equal to $p_S/2$ or just below $p_S/2$. Yet, if the recipient not only notes whether $x_2 \leq p_S/2$, but also takes into account the exact amount of x_2, then $x_2 = p_S/2$ may make her 'particularly distrustful'. Therefore, an extended analysis should show that all types of dictators would not necessarily propose an amount of approximately $x_2 = p_S/2$ but that every type may have a personal optimum (depending on the distribution of types which determines the updating of the recipient's beliefs). Of course, a dictator can hide his large pie only if there are other dictators who would offer the same amount x_2 when endowed with a small pie.

The third type of dictator (endowed with p_L) would offer $x_2 < p_S/2$ even under complete information. He will further reduce x_2. Then, there is the fourth type of dictator who does not give anything under complete information. He does not change his strategy with a change of information status. Hence in comparing Dictator Game experiments with complete and incomplete information, the following deductions can be made: the income-based approach predicts no differences. The utility-based approach predicts four different behavioural patterns for dictators endowed with p_L. First, there are dictators who offer a larger share up to the equal split under complete information and who are expected not to change their behaviour under

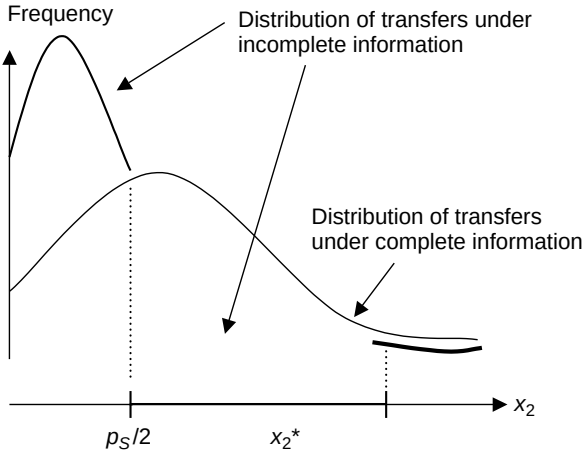


Figure 9.1 Expected changes of the distribution of transfers from a dictator endowed with m_L

incomplete information. The second type are those dictators who do not give much more than $p_S/2$ and the third type are those who give less than $p_S/2$ under complete information. Under incomplete information the second type will propose less (but close to) $p_S/2$, the third type will further reduce their offer, and the fourth type will transfer nothing under both conditions (see Figure 9.1, bold curve).

3 Experimental design and procedure

The present experiment encompassed two different treatments, the Standard Dictator Game (game 1) and the Dictator Game under asymmetric information (game 2). The Standard Dictator Game is aimed at confirming previous results and serving as a baseline for comparison with game 2. In both treatments, a dictator was anonymously matched with a recipient. The dictator received an endowment of either 10 euros (the large pie) or 1.15 euros (the small pie).[5] In game 1 the recipient knew the endowment of the dictator. Game 2 differed from game 1 only in that the recipient did not know the exact size of the endowment but was informed that the dictator received a large pie with probability 2/3 and a small pie with probability 1/3. All other variables were kept constant in both games (for the instructions see the Appendix).

Two hundred and forty undergraduates from the European University Viadrina participated in the experiment – 120 in each session. They were recruited through announcements in lectures. Participation required appearance at a prearranged place and time and was restricted to one

session. Upon arrival, participants were randomly assigned their roles as dictators (person A) or recipients (person B). In both treatments 40 dictators were endowed with the large and 20 with the small pie. Throughout the sessions participants were placed in two separate rooms. All experiments were conducted once, after the participants had received written and verbal instructions about the setting. All participants were randomly and anonymously matched.

In each session all dictators received an envelope containing the written instructions and the amount of the pie, split into many coins (in German Pfennig), enabling the dictator to propose any amount he preferred to the recipient. The written instructions contained the same script for both treatments, the only modification being that in game 1 the recipient was informed about the size of the pie the dictator had received, and in game 2 the recipient was informed about the probability of the dictator having received the one or the other pie. To ensure complete privacy for the decision, cubicles were offered. The dictators put the amount given to the recipient back into the envelope, and put the envelope into a box where all proposals were collected. They pocketed their own share of the pie. The box was then transferred to room B and randomly distributed to the recipients after two neutral persons had registered the amount in each envelope in a third room. Thus, it was not possible to match any individual action to any particular subject.

When the experiment was designed it became clear that the smaller the small pie in the incomplete information setting, the easier it became to discriminate between the behaviour of those dictators in the two treatments who received the large pie. At the same time, the smaller the small pie, the less it became possible to discriminate between the behaviour of those dictators in the two treatments who received the small pie. Since the decisions of dictators who received the large pie were our central focus, we chose the small pie to be 'very small' – €1.15. Accordingly, we will restrict the analysis to those dictators who received the large pie.

4 Predictions

Starting with the unique 'egoistic' equilibrium prediction, the dictator would make no positive offer to the recipient no matter what information the recipient is given about the dictator's pie size (H0a). Applying the income-based approach to altruism is straightforward. It results in (H0b); the distribution of the dictator's proposals should be the same under both conditions, irrespective of the information status of the recipient.

Applying the utility-based approach to altruism leads to the following hypotheses. In comparison with game 1, among those who received the large pie in game 2 (where the recipient has incomplete information),

we expect that:

Hypothesis (H1) more subjects will offer nothing or less than 0.6 euro (around half the small pie);
Hypothesis (H2) less subjects will offer amounts between 0.6 euro and 2.6 euro; and
Hypothesis (H3) about the same number of subjects will offer between 2.6 and 5 euro.[6]

This leads to the overall hypothesis that dictators who are endowed with a large pie will, on average, make lower proposals under incomplete information than under complete information.

5 Experimental results

To begin with, we examined whether the data of the present €10 Dictator Game under complete information was similar to the $10 Dictator Game experiment of Forsythe *et al*. (1994: 366) – which served as a baseline treatment in previous studies. A comparison shows that the distribution of proposals are similar: in the modal offer most of the participants offer between 20 and 25 per cent of the pie, and in the average the payoff is 22.3 per cent in Forsythe *et al*. and 20.4 per cent in the present experiment (no significant difference in behaviour is found by the Mann–Whitney U-test, $p = 0.2912$). This indicates that the behaviour of the present 'population' is comparable with earlier observations.

The present study focuses on a comparison of the dictator's willingness to transfer a certain share of his pie to the recipient when the information status of the recipient is variable. Starting with the 'overall hypothesis' of the average offer, dictators who had received the €10 pie in game 2 proposed on average 11.4 per cent of the pie, half of what dictators offered under complete information in game 1 (20.4 per cent) (an overview of all offers in the two games is given in Table 9.1). A Kolmogorov–Smirnov test verifies that the distribution of proposals is significantly lower in game 1 than in game 2 ($p < 0.01$). H0a and H0b can be rejected.

This leads directly to the three hypotheses H1 to H3. In Figure 9.2 we have ordered the results of games 1 and 2 in such a way that the hypotheses can be compared. Starting with the share of participants who made proposals close to the unique 'egoistic' equilibrium outcome, that is nothing or less than 60 per cent, we observe an increase from 23 per cent to 60 per cent in the asymmetric information game (in support of H1, Fisher's probability test shows $p = 0.017$).

This observation indicates that some participants did try to signal to the recipient that they had received the small rather than the large pie. Moreover, the high share of participants giving even less than half the small pie further

Table 9.1 Offers of dictators in game 1 (complete information) and game 2 (incomplete information)

Dictator offer in	*Game* 1	*Game* 2
0.00	4	11
0.10	–	3
0.16	1	–
0.26	–	2
0.31	1	–
0.42	–	1
0.52	1	2
0.57	2	5
0.62	2	1
0.94	–	1
1.00	1	–
1.04	–	1
1.15	6	4
1.67	3	–
2.08	1	–
2.19	6	–
2.34	–	1
2.59	1	–
2.66	–	2
2.92	–	1
3.23	2	–
3.65	1	–
3.75	–	1
4.00	1	–
4.27	1	–
4.48	–	1
4.53	1	–
4.74	–	1
4.80	2	2
5.00	3	–

supports our approach in two ways (see also the distribution in Figure 9.2): not only did participants who had given less than half of the small pie in game 1 reduce their offers further to (almost) zero, but even participants who had given slightly more than $p_S/2$ reduced their proposals to an individually calculated optimum which was not necessarily equal to $p_S/2$. This leads us to (H2).

Figure 9.2 also provides answers to the two further hypotheses: In (H2) we hypothesized a sharp decrease of offers in the range between 0.60 and 2.60 euros since it was expected that those dictators who gave less than $p_S/2$ in game 2 would have been of this type. Our data suggest this to be true since there were only 20 per cent in game 2 – as opposed to 50 per cent

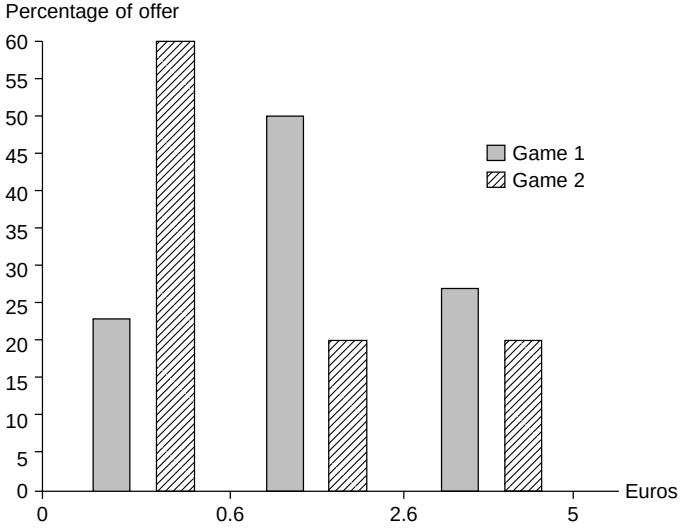


Figure 9.2 Cumulative results of dictator transfers in game 1 and game 2 (according to hypotheses 1 to 4)

in game 1 – who offered an amount between 60 cents and 2.60 euros (in support of H2, $p = 0.005$).

Coming to the final hypothesis (H3) it was asserted that dictators who transfer a relatively high share of the large pie under complete information would do the same thing under incomplete information. The reason for this expectation is that this type of dictator might be ready to signal to the recipient that he has a large pie. Since he is willing to sacrifice more than the average offer in favour of the recipient he probably expects the recipient to be satisfied with the proposal and therefore make the proposal utility increasing. The share of dictators of this type was 27 per cent in game 1 and 20 per cent in game 2, showing no significant difference ($p = 0.3$) between the settings.

Last but not least, some remarks are needed about the division of the small pie under incomplete information. As mentioned above, due to the small size of the small pie we did not expect any significant differences in behaviour between persons in the two settings and the experimental results showed that the decisions were indeed very similar. However, it is interesting to note that all dictators endowed with a small pie in the incomplete information game transferred either nothing or less than €0.6. Thus, it was possible for those persons A who were endowed with a large pie and who gave less than €0.6 (60 per cent of the participants with the large pie gave less than €0.6 in the incomplete information game), not to reveal the true size of their pie, since all participants endowed with a small pie gave a similar amount to person B. (This argument is, of course, based on rational expectations of the subjects.)

The outcome of the experiment provides support in favour of hypotheses (H1), (H2) and (H3) and rejects the hypotheses (H0a) and (H0b).

6 Summary

The present experiment compares the willingness of dictators to make offers to anonymous recipients when the information status of the recipient is variable. In the baseline treatment where the recipient is fully informed about the pie size, dictators give similar amounts as in previous studies. In the second treatment where the recipient is informed only about the probability distribution of the two pie sizes, dictators still give non-trivial amounts, but some of them significantly reduce their transfers.

In the treatment with incomplete information about the two pie sizes we differentiate between four types of dictators (among those dictators who received a large pie): dictators who keep the complete pie for themselves; dictators who transfer less than the small pie; dictators who transfer a little more than half of the small pie; and dictators who offer an equal split of the large pie. Having received the large pie, the first and the fourth type of dictators did not change their behaviour under incomplete information. The intermediate types, however, instead of revealing their pie size, preferred to hide the amount of their own income by reducing the offer. They tried to cause the recipient to believe that she had received a considerable amount of the small pie instead of a small amount of the large pie.

Utility-based approaches to altruism (which are more favoured in psychology than in economics) are able to give a thorough explanation of the observed behaviour if, in addition, we introduce a fairness component in the utility functions. Then, under incomplete information a dictator often has the choice to increase the recipient's utility by offering more money or to pretend to be more fair (which in this case is associated with giving less money). Certain types of dictator who are ready to transfer substantial amounts face this problem (and decide differently). For the type of dictators who transfer rather small amounts or no money, there is no question of pretending to be fair.

The joint hypotheses of utility-based and fairness-shaped altruism is capable of explaining the results of a dictator game under incomplete information. We think that this approach should be pursued in theory as well as for explaining the results of other experiments.[7]

Appendix: instructions to the players in the dictator experiment

In the description the instructions for player A are presented. Differences corresponding to the two treatments are indicated in boldface. For the instructions of person B the obvious changes were made.

Instructions for player A

You have been asked to participate in an economics experiment on individual decision-making. For your participation you may earn some money which will be paid to you right away. Before you make any decision please read the following instructions carefully. If you have any questions, do not hesitate to ask the experimenter.

In this experiment each of you will be paired with a different person who is in another room. This is room A and you are person A. The person who will be paired with you is person B in room B. You will not be told who these people in room B are either during or after the experiment, and they will not be told who you in room A are either during or after the experiment. You will notice that there are other people in the same room with you who are also participating in the experiment. You will not be paired with any of these people. The decisions that they make will have absolutely no effect on you nor will any of your decisions affect them.

Game 1 is conducted as follows: A sum of DM 19.55 (DM 2.25) has been allocated to you in coins in the envelope. The person B who is matched with you knows that you have received this amount. You are now asked to propose how much of this each person is to receive. You are free to propose any amount you like to give to person B: nothing, something or the whole sum.

Game 2 is conducted as follows: A sum of DM 19.55 (DM 2.25) has been allocated to you in coins in the envelope. There are 39 (40) more players who received DM 19.55 and 20 (19) more players who received DM 2.25. Person B who is paired with you does not know the exact amount allocated to you. Person B knows that you received DM 2.25 with a probability of 33.3 per cent and DM 19.55 with a probability of 66.7 per cent. You are asked to propose how much of the amount of DM 19.55 (DM 2.25) each person is to receive. You are free to propose any amount you like to give to person B: nothing, something or the whole sum.

For your decision you may use the cubicles in the room. You will have five minutes to come to a decision about your proposal. If you made your decision about the amount which you like to propose to person B, put the respective amount into the envelope and put the envelope into the box next to your cubicle. Then you may pocket the amount you have allocated to yourself right away. Do not talk to the other people in your room until your session is completed. Do not be concerned if other people make their decision before you.

Notes

1 See, for example, Simon (1957), Becker (1974) and Collard (1978) for early analysis from an economic perspective and Homans (1961), Bryan and Test (1967) and Rosenhan and White (1967) for analysis from a psychological perspective.
2 Recent research shows that altruism may be evolutionary stable, at least under certain conditions (see Bester and Güth, 1999; Bolle, 2000).
3 Güth and Huck (1997) conducted a similar experiment where the recipient had incomplete information about the pie size. In contrast to the present experiment, they used the strategy method and had no control setting under complete information, which made it impossible to use their data for our test. It should also be emphasized that the focus of their paper was completely different from ours.
4 Recent theoretical approaches of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) also made use of psychological equity theories. However, these approaches are not appropriate for explaining the behaviour in the Dictator Game experiment under incomplete information, since they rely on the income-based approach of

equation (1). For further discussions of these approaches from our point of view, see Kritikos and Bolle (2001).

5 The experiment was conducted in November 2001 using German currency. Thus, the pie size was then 19.55 DM for the large and 2.25 DM for the small pie. Hence, no prominence effects could be found in the experiment.

6 The experiment was conducted in German currency where 2.6 euros transfers to DM 5.

7 A first example of similar behaviour outside the laboratory relates to a situation when the bequest of two children needs to be determined. Stark and Zhang (2002: 21) argue: 'Parents who are equally altruistic towards their children may consider leaving a larger bequest to the lower-earning child 2 (a "compensation" act). However, because the division of bequests is public information, unequal division is tantamount to a public statement that child 2's earnings are relatively low – a declaration that can embarrass child 2.' This argument is in line with the utility-based approach to altruism.

References

Andreoni, J. and J.H. Miller (2002) 'Giving According to GARP: An Experimental Study of Rationality and Altruism', *Econometrica*, vol. 70, pp. 737–53.

Andreoni, J. and L. Vesterlund (2001) 'Which Is the Fair Sex? Gender Differences in Altruism', *Quarterly Journal of Economics*, vol. 116, pp. 293–312.

Arrow, K.J. (1975) 'Gifts and Exchanges', in E.S. Phelps (ed.), *Altruism, Morality and Economic Theory* (New York: Russell Sage Foundation), pp. 13–28.

Barro, R.J. (1974) 'Are Government Bonds Net Wealth?', *Journal of Political Economy*, vol. 82, pp. 1095–117.

Becker, G.S. (1974) 'A Theory of Social Interactions', *Journal of Political Economy*, vol. 82, pp. 1064–93.

Bester, H. and W. Gueth (1998) 'Is Altruism Evolutionarily Stable?' *Journal of Economic Behavior and Organization*, vol. 34, pp. 193–209.

Bolle, F. (1991) 'On Love and Altruism', *Rationality and Society*, vol. 3, pp. 197–214.

Bolle, F. (2000) 'Is Altruism Evolutionarily Stable? An Envy and Malevolence? Remarks on Bester and Güth', *Journal of Economic Behavior and Organization*, vol. 42, pp. 131–3.

Bolle, F. and A.S. Kritikos (2004) 'Altruistic Behavior under Incomplete Information', Discussion Paper, European University Viadrina, Frankfurt (Oder).

Bolton, G.E. and A. Ockenfels (2000) 'A Theory of Equity, Reciprocity and Competition', *American Economic Review*, vol. 90, pp. 166–93.

Bolton, G.E. and R. Zwick (1995) 'Anonymity versus Punishment in Ultimatum Bargaining', *Games and Economic Behavior*, vol. 10, pp. 95–121.

Bryan, J.H. and M. Test (1967) 'Models and Helping: Naturalistic Studies in Helping Behavior', *Journal of Personality and Social Psychology*, vol. 6, pp. 400–7.

Camerer, C.F. and R.H. Thaler (1995) 'Ultimatums, Dictators and Manners', *Journal of Economic Perspectives*, vol. 9, pp. 209–19.

Collard, D. (1978) *Altruism and Economy* (Oxford: Martin Robertson).

Eckel, C.C. and P. Grossman (1996) 'Altruism in Anonymous Dictator Games', *Games and Economic Behavior*, vol. 57(16), pp. 181–91.

Fehr, E. and K. Schmidt (1999) 'A Theory of Fairness, Competition and Cooperation', *Quarterly Journal of Economics*, vol. 114, pp. 817–68.

Forsythe, R.J., L. Horowith, N.E. Savin and M. Sefton (1994) 'Fairness in Simple Bargaining Experiments', *Games and Economic Behavior*, vol. 6, pp. 347–69.

Güth, W. and S. Huck (1997) 'From Ultimatum Bargaining to Dictatorship – An Experimental Study of Four Games Varying in Veto Power', *Metroeconomica*, vol. 48, pp. 262–79.
Hannson, R.O. and K.M. Slade (1977) 'Altruism towards a Deviant in City and Small Town', *Journal of Applied Social Psychology*, vol. 7, pp. 272–9.
Hoffman, E., K. McCabe, K. Shachat and V. Smith (1994) 'Preferences, Property Rights and Anonymity in Bargaining Games', *Games and Economic Behavior*, vol. 7, pp. 346–80.
Homans, G.C. (1961) *Social Behavior. Its Elementary Forms* (New York: Harcourt).
Konow, J. (2001) 'Fair and Square: the Four Sides of Distributive Justice', *Journal of Economic Behavior and Organization*, vol. 46, pp. 137–64.
Kritikos, A.S. and F. Bolle (2001) 'Distributional Concerns: Equity or Efficiency Oriented?', *Economics Letters*, vol. 73, pp. 333–8.
Mikula, G. (1983) 'Justice and Fairness in Interpersonal Relations: Thoughts and Suggestions', in H. Tajfel (ed.), *The Social Dimension* (Cambridge: Cambridge University Press), pp. 204–27.
Latané, B. and J. Rodin (1969) 'A Lady in Distress: Inhibiting Effects of Friends and Strangers on Bystander Intervention', *Journal of Experimental Social Psychology*, vol. 5, pp. 189–202.
Levine, D. (1998) 'Modelling Altruism and Spitefulness in Experiments', *Review of Economic Dynamics*, vol. 1, pp. 593–622.
Reis, H.T. (1983) 'The Multidimensionality of Justice', in R. Folger (ed.), *The Sense of Injustice: Social Psychological Perspectives* (New York: Plenum), pp. 25–61.
Rosenhan, D. and G.M. White (1967) 'Observation and Rehearsal as Determinants of Pro-social Behavior', *Journal of Personality and Social Psychology*, vol. 5, pp. 424–31.
Simon, H.A. (1957) *Models of Man* (New York: Wiley).
Simon, H.A. (1993) 'Altruism and Economics', *American Economic Review*, vol. 83, pp. 156–61.
Stark, O. and J. Zhang (2002) 'Counter-compensatory Inter-vivos Transfers and Parental Altruism: Compatibility or Orthogonality?' *Journal of Economic Behavior and Organization*, vol. 47, pp. 19–25.
Walster, E., G. Walster and E. Berscheid (1978) *Equity: Theory and Research* (Boston: Allyn & Bacon).
Wilson, J.P. (1976) 'Motivation, Modeling, and Altruism: A Person × Situation Analysis', *Journal of Personality and Social Psychology*, vol. 34, pp. 1078–86.

10

Equity Judgements Elicited Through Experiments: An Econometric Examination*

Jochen Jungeilges
University of Bielefeld, Germany
and
Theis Theisen
Agder University College, Norway

1 Introduction

Utilitarianism and Rawlsianism stand out as prominent schools of social welfare assessment. According to the utilitarian school, welfare judgements should be based on how policies affect the sum of individual utilities. By contrast, according to the *maximin* principle of Rawls (1971), the Rawlsian school claims that welfare judgements should be based on how policies affect the utility of the worst-off individual in society.

In a nutshell, the Rawlsian *maximin* principle can be illustrated as follows. Imagine a society of two individuals, 1 and 2. This society is considering two policies, x and y. Individual 1 prefers x to y, while individual 2 prefers y to x. Assume that individual 2 is always better off than individual 1, irrespective of the policy chosen. According to the Rawlsian equity principle underlying the *maximin* principle, policy x should then always be socially preferred to y. In this paper, we examine whether individual decisions are consistent with the Rawlsian equity principle. Economic theory cannot provide an answer to this question. Hence our examination will be empirical.

* The authors wish to thank Wulf Gaertner, Geir Asheim, Lars Schwettman and Oded Stark for helpful comments. We are also indebted to seminar participants at the Department of Economics and Business Administration at Agder University College, at the Department of Economics at Osnabrück University and to participants at the XIII World Congress of the International Economic Association. Much of the research was carried out at the Department of Economics at Bielefeld University. In particular, we thank Volker Böhm for his hospitality and for providing us with excellent working conditions. *Norsk Faglitterær Forfatterforening* is gratefully acknowledged for providing Theis Theisen with a travel grant.

Empirical examinations, however, are afflicted with their own problems. In particular, actual choices will usually be determined by a mix of ethical and selfish considerations, the constraints under which choices are made, and strategic considerations. Thus, it may be difficult to recover the underlying ethical principles from observed choices. A number of economists have therefore turned to experiments as a method for eliciting the principles that guide individuals when prioritizing on behalf of society. As pointed out by Hargreaves-Heap and Hollis (1987), the experimental approach does not escape the typical philosophical criticism of empiricism. Nevertheless, it has been demonstrated in previous research that experiments can reveal principles that guide choices.

Yaari and Bar-Hillel (1984) carried out one of the first experiments aimed at revealing the principles that guide individuals who are given the hypothetical task of allocating goods to others. After having examined nine different principles, including the Rawlsian *maximin* principle and Utilitarianism, they concluded that people in experiments tended to act in accordance with the Rawlsian *maximin* rule when taking decisions in situations involving 'need'. In other situations, however, they found that the *maximin* rule did not apply. Frohlich, Oppenheimer and Eavey (1987a) and Frohlich, Oppenheimer and Eavey (1987b) found that the vast majority of participants preferred a compromise between the Rawlsian *maximin* principle and Utilitarianism, rather than one of these 'extreme' principles. In a third experiment, reported by Gaertner (1994), the results are again mixed and difficult to understand as the outcome of one of the pure principles set out in the literature.

When it comes to an experimental setup, the present paper is in line with the papers of Gaertner, Jungeilges and Neck (2001), and Gaertner and Jungeilges (2002). Gaertner, Jungeilges and Neck (2001) summarize the state of the literature and conclude that whether or not people base their welfare assessments on the Rawlsian equity principle is dependent both on context and on the political and cultural environment. It seems, however, that the empirical support for pure Utilitarianism is rather weak.

Sound empirical evidence requires statistical rigour. As emphasized by Manski (2002), data generated in economic experiments are often not subjected to state-of-the-art statistical analysis. For instance, even in the papers of Frohlich, Oppenheimer and Eavey (1987a, 1987b) and Gaertner and Jungeilges (2002), who do carry out some formal statistical testing, it seems that the statistical analysis could be carried further.

The reward for doing so may be substantial. First, without exploiting statistical techniques there is a risk of invalid inference. In addition, formulating econometric models that relate possible outcomes of an experiment to variables characterizing probants, the society in which probants live, and the context of choice, may help to unravel why different individuals facing the same decision problem may make different choices, and why an

individual's decision may be context dependent. This strategy is followed in the present chapter. The emphasis on econometric modelling distinguishes this contribution from the papers mentioned above. Our data were generated through an experiment in which Norwegian students were subjected to an established experimental design developed by Gaertner (1992). We find that the set of individual characteristics that matters for choice differs across contexts. Two factors, *gender* and certain aspects of *education*, seem to play a crucial role in most of the decision situations considered.

In section 2, we provide a brief account of the social choice context motivating the experiment. A discussion of the experimental design, and the 'instrument' used in the empirical examination is given in section 3. The data-collection procedure and the main features of probants are highlighted in section 4. The subsequent section contains results on an individual's propensity to act in accordance with the Rawlsian equity principle. Moreover, we test whether these results depend on context, education and gender. In section 6, a binary response model is used to explore factors that drive individuals to act in accordance with the Rawlsian equity principle. In addition, we rely on the same class of models to examine possible reasons for departing from this Rawlsian position. Section 7 summarizes the main results and outlines some ideas for further research.

2 Theoretical background

Let $X = \{x, y, \ldots\}$ represent a finite or infinite set of social states. The set $N = \{1, 2, \ldots, i, j, \ldots, n\}$ refers to a finite group of individuals. Suppose that $X \geq 3$ and $N \geq 3$. Next, define $\mathcal{R}$ as the set of orderings on the set of social states X. Then for $R \in \mathcal{R}\ \forall x, y \in X$, we write xRy to indicate that a social state (or policy) x is at least as good as the state y from a collective point of view.

To reflect the evaluations of social states by individuals, consider the Cartesian product of the set of individuals and the set of social states: $X \times N$. Elements of this set are of the form (x, i). Such pairs are interpreted as referring to person i under social state x. Let $\mathcal{U}$ denote the set of bounded functions defined on $X \times N$. Functions from this set are used for welfare comparisons between individuals under a given social state as well as between different individuals across alternative social states. Given $u \in \mathcal{U}\ \forall i, j \in N$ and $\forall x, y \in X$, the statement $u(x, i) \geq u(y, i)$ says that social state x is at least as good as social state y from the point of view of individual i. To express that under social state x individual i is at least as well off as as individual j under social state y, we write $u(x, i) \geq u(y, j)$. Social choice theory is concerned with finding characterizations of social welfare functionals. That is, one tries to characterize the type of functions which can be defined from the set $\mathcal{U}$ to the set of orderings $\mathcal{R}$, given some reasonable restrictions. Among the typical requirements such

as independence of irrelevant alternatives, Pareto-type principles and the anonymity principle one finds Hammond's *equity axiom*:

Axiom 1 For some $u \in \mathcal{U}$ and any $x, y \in X$, if for some pair of individuals $i, j \in N$

$$u(y, i) < (x, i) < u(x, j) < u(y, j)$$

and for all $k \in N \backslash \{i, j\}$: $u(y, k) = u(x, k)$, then xRy

This axiom says that the individual who is better off should not determine the social ordering. The relationship with Rawls's second principle mentioned in the introduction is apparent. At this point it is sufficient to state that this *equity axiom* is an important one. Its technical significance and extention are discussed in Deschamps and Gevers (1978) and Gaertner (1992). We now turn to the experimental design that we will use to examine whether human decisions are consistent with Axiom 1.

3 Experimental design

In order to elicit peoples' attitudes towards the type of distributional issues described in the preceding section, six different situations or decision contexts were used. This experimental design, developed by Gaertner (1992), has been used in experiments at several universities in various countries. The format of all situations is of the same type as the general format described in section 2. In addition to a copy of the material administered to probants provided in the Appendix, the main characteristics of the situations are given in Table 10.1.

In each situation, there is always one group (person) that is always better off under x than under y, while the other group (person) is better off under y than under x. Situation 1 is the most transparent and can be used to explain the format more explicitly. In that situation, in the baseline question probants are asked to decide whether a certain amount of money should be allocated exclusively to assist a handicapped person (alternative x), or to educate an intelligent child (alternative y). In the second question, alternative x is still to allocate the money to assist the handicapped person, but the number of intelligent children that could get education in alternative y is increased to two. In the third and the fourth (and last) steps, the number of intelligent children who would benefit from alternative y is increased to three and four, respectively. In other words, more and more individuals who unanimously would prefer alternative y to x are gradually introduced, while the number of individuals who would benefit from x is kept constant at one. The crucial question is whether the number of individuals who would benefit

Table 10.1 Characteristics of situations

Situation	*Fund*	*Alternative x*	*Alternative y*
1	Unspecified amount	1 disabled individual trained to master basic tasks	Improve language and science skills of (o) 1 gifted child (a) 2 gifted children (b) 3 gifted children (c) 4 gifted children
2	Central bank profit	Food aid to Sub-Saharan Africa	Environmental projects to (o) improve ecological conditions in the coastal area (a) and reduce air pollution of coal-fired power plants (b) and clean rivers programme (c) and reduce noise along highways
3	Exchange reserve	Purchase dialysis equipment	Purchase vitamins and fruit to upgrade diets of (o) pregnant women (a) and infants (b) and teenagers (c) and workers doing hard physical labour
4	Exchange reserve	Purchase dialysis equipment	Import of wine affordable for (o) well-off citizens (a) and less well-to-do citizens (b) and even less prosperous (c) and even less prosperous
5	Exchange reserve	Import cheap textiles to benefit the needy	Import of wine affordable for (o) well-off citizens (a) and less well-to-do citizens (b) and even less prosperous (c) and even less prosperous
6	Conditional reconstruction loan	Grant *right to strike* and *freedom of occupational choice* to benefit the country's workers	Eliminate *right to strike* and *freedom of occupational choice* to benefit (o) employees of large enterprises (a) and self-employed with small or medium-sized businesses (b) and civil servants (c) and retired persons

from alternative y will affect the allocation of money, when money must be allocated either exclusively to x or exclusively to y.

In situations 2, 3, 4, 5 and 6, the alternatives among which probants are asked to prioritize differ from situation 1. In situation 2, probants are asked whether they would allocate money to an aid programme against hunger in Sub-Saharan Africa (x) or to environmental protection programmes in or close to the probants' home country. In situation 3 the issue at stake is whether a poor country should allocate its limited amounts of foreign currency to the purchase of dialysis machines (x) or to the purchase of fruit and vitamin pills (y) to selected parts of the population. In situation 4, the issue is whether a poor country should allocate foreign currency to the purchase of dialysis machines (x) or to the import of Bordeaux wine (y). In situation 5, the question is whether a poor country should allocate foreign currency to the purchase of clothing for a group of needy people (x) or to the import of Bordeaux wine (y). Finally, in situation 6 the issue is whether a run-down country should emphasize workers' rights to strike and to choose occupation freely and pull itself up by its bootstraps (x) or accept a condition to set aside workers' basic rights in order to obtain a favourable loan that would benefit some or all of the following groups: large enterprises, the self-employed with small or medium-sized firms, civil servants and retired persons.

All situations are designed so that there will be unanimous support for the claim that alternative x is a social goal worth pursuing. In situations 1, 2, 3 and 6, the same is likely to hold for alternative y. Situations 4 and 5, however, were deliberately designed so that some may consider alternative y not to be warranted as a social goal. This was done in order to test the logic and consistency of probants' answers across situations and to test for context dependence of choices.

It is natural to interpret situation 1 as referring to the probant's own country, with the money available for policies x and y being government revenues. Moreover, the handicapped person that would benefit from policy x could well be the probant herself, after let us say a car accident, and the intelligent child that could benefit from policy y could well be the still unborn child of the probant. Consequently, situation 1 comes close to the context behind the Rawlsian concept of the veil of ignorance and Harsanyi's choice under uncertainty.

Situations 3, 4, 5 and 6 differ from 1 in the sense that they clearly relate to other countries than the one where the probants live. In these situations, it is clear that the probants will neither have to contribute to the funding of policies x and y, nor will they benefit from those policies. Consequently, in situations 3, 4, 5 and 6, probants play the role of an outside judge. When it comes to situation 2, it is less clear how closely related it is to the probants, but it is reasonable to interpret it as lying somewhere between situation 1 on the one hand and situations 3, 4, 5 and 6 on the other hand.

In addition to questions about how given resources should be allocated, probants were asked about their demographic characteristics, parental professional background and expected position in a future income distribution. Such variables may serve several purposes. First, they may be used to assess whether the sample is representative. Second, they may be used as explanatory variables in an econometric examination of what drives probants to make different distributional choices.

4 Data collection and features of the sample

The data were collected in March 2001, using two different student groups at Agder University College located on the southern coast of Norway. Group 1 consists of first-year students in a two-year study programme in basic business administration (BBA). Group 2 consists of third-year students in an advanced two-year programme in businesss administration (ABA). In order to be admitted to the ABA-programme, it is required that students have passed the exams in the BBA-programme, either at Agder University College, or at another Norwegian college. Only students with high grades from the BBA-programme were admitted to the ABA-programme.

Descriptions of the six situations together with answering sheets were administered during normal lecture hours. It was pointed out to the students that there was no such thing as a single right answer to a question. In order to give students an incentive to complete the questionnaires, they were informed that a winner would be drawn randomly from those who handed in completed forms, and that the winner would be given a canteen card worth 150 Nkr (€19).

The first-year students had not been exposed to welfare theory or social choice theory prior to the experiment. The third-year students, by contrast, had been introduced to the concepts of utilitarian and Rawlsian welfare functions. Discussions with the students after the answering sheets had been collected revealed that at least some of the third-year students had worked out that the experiment was related to welfare economic issues and the concepts of Rawlsianism and Utilitarianism.

In the class of first-year students, 85 students were present when descriptions of the six situations and answering sheets were handed out, but 12 students left the lecture hall without completing the answering sheets. In addition, 7 of the forms handed in were too incomplete to be used. This leaves us with a sample of 66 first-year students. In the class of third-year students, 70 were present when the descriptions and answering sheets were handed out. Only 3 of them left without completing their forms. In addition, 3 forms which were not filled out completely had to be discarded. This left a sample of 64 third-year students. Hence, we were left with a total sample size of 130. Of these, 20 did not provide any information on demographic and other background variables. Therefore, in analyses relating decisions to

demographic and background information, we have to rely on the sample of 110 complete answers.

Of the 110 respondents who provided complete information about demographic and other background variables, there was an equal split between males and females. Probants' age ranged from 19 to 40, but 95 per cent of them were between 19 and 25 years of age. About one-third had job experience. Probants identified themselves as having parents who were workers (7%), craftsmen (13%), skilled workers (4%), employed in public administration (25%), employed in the private sector (39%) or self-employed (12%). A comparison of these numbers with official statistics indicates that the sample is not too far from being representative when it comes to social background, although worker background is under-represented and self-employed background over-represented. When it comes to future expectations, however, probants did not consider themselves to be average individuals. This is seen from the fact that 40 per cent expected to earn an income in the upper (fourth) quartile of the income distribution, 50 per cent expected to earn an income in the third quartile, but only 10 per cent expected to earn an income that would put them in the lowest two quartiles of the income distribution.

5 Data analysis

For each situation, four decisions – indexed by $j \in \{1, \ldots, 4\}$ – constitute the response of a probant. A decision in favour of the individual(s) who is (are) the worst off under both policies is coded as 0, while the alternative decision is recorded as 1. For each situation, a probant's response is therefore represented as a binary string consisting of four digits. The string 0011, for example, indicates that a probant decided in line with the equity axiom in the baseline question. She maintained the Rawlsian position under the first increase of the group of those who are better off anyway. However, she left the Rawlsian position as that group was increased further in question $j = 3$.

There are 2^4 ways in which two distinct objects, 0 and 1, can be arranged in four places. The set of 16 possible outcomes represents the elementary events of the experiment under any of the decision situations. The 16 binary strings have a decimal representation: $0, 1, 2, \ldots, 15$, cf. columns 1 and 2 of Table 10.2. These integers can be viewed as the possible realizations of a univariate discrete random variable D.

The logic behind the event 0000 is Rawlsian, while it is non-Rawlsian for element 1111. The three elements (0111, 0011, 0001) are in accordance with Rawlsian logic in the baseline question ($j = 1$). It is however the case for these elements that from question j on, $j \in \{2, 3, 4\}$, responses are not consistent with Rawlsian logic. To avoid ambiguities later on, let us define the subset C of events that are consistent with either the Rawlsian or the non-Rawlsian logic:

Definition 1 $C = \{0000, 0001, 0011, 0111, 1111\}$

Table 10.2 Estimated densities for all situations ($N = 130$)

Cases	*Dec.*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0	0.492	0.492	0.423	0.915	0.815	0.462
0 0 0 1	1	0.038	0.077	0.046	0.023	0.085	0.077
0 0 1 0	2	0.008	0.000	0.008	0.000	0.000	0.077
0 0 1 1	3	0.123	0.100	0.146	0.046	0.062	0.169
0 1 0 0	4	0.000	0.015	0.008	0.000	0.008	0.015
0 1 0 1	5	0.000	0.000	0.008	0.000	0.000	0.000
0 1 1 0	6	0.015	0.015	0.015	0.000	0.000	0.031
0 1 1 1	7	0.192	0.054	0.154	0.008	0.008	0.038
1 0 0 0	8	0.015	0.038	0.008	0.008	0.000	0.008
1 0 0 1	9	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10	0.000	0.023	0.000	0.000	0.000	0.000
1 0 1 1	11	0.000	0.008	0.000	0.000	0.000	0.000
1 1 0 0	12	0.000	0.008	0.015	0.000	0.008	0.000
1 1 0 1	13	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14	0.000	0.023	0.015	0.000	0.000	0.008
1 1 1 1	15	0.115	0.146	0.154	0.000	0.015	0.192

The 11 events in the complement C fall into two categories. The first category consists of the eight events with decimal representations 2, 4, 5, 6, 9, 10, 11 and 13, where the individual switches back and forth between deciding in accordance with the equity axiom and not adhering to the axiom. Such decision patterns are clearly inconsistent and may indicate that the individual has not understood the logic of the experiment. In the second category consisting of the three events 8,12 and 14, the individual at an early stage does not decide in accordance with the equity axiom, but at a later stage (s)he does adhere to that axiom. Such decision patterns seem peculiar. Hence, all events in C are referred to as *inconsistent* throughout the rest of the paper.

Decision patterns

In this section, we report density estimates for the six situations. Let us first consider the probability estimates for all possible decision patterns. The point estimates given below are based on all records in the sample. The associated 95 per cent confidence intervals for the relative frequencies $\hat{p}$ are indicated by brackets in Figures 10.1 to 10.6. For details concerning the computation of the intervals see Hogg and Craig (1978: 215–17).

The estimated densities for situations 4 and 5 are clearly different from the estimates obtained for the remaining situations. In situation 4, as well as in situation 5, we find a high relative frequency of support for the worst-off individual(s), no matter how large the set of individuals who are better off under any alternative. This tendency is most pronounced in the case of situation 4, where 92 per cent of the probants favour the worst-off individuals (compared

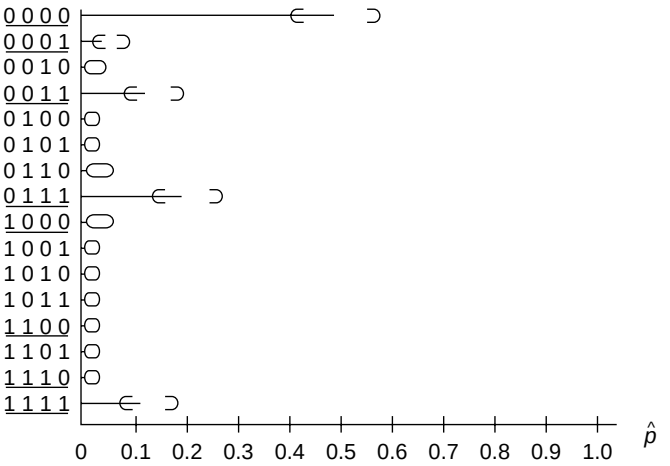


Figure 10.1 Interval estimates of probabilites for situation 1 ($N = 130$)

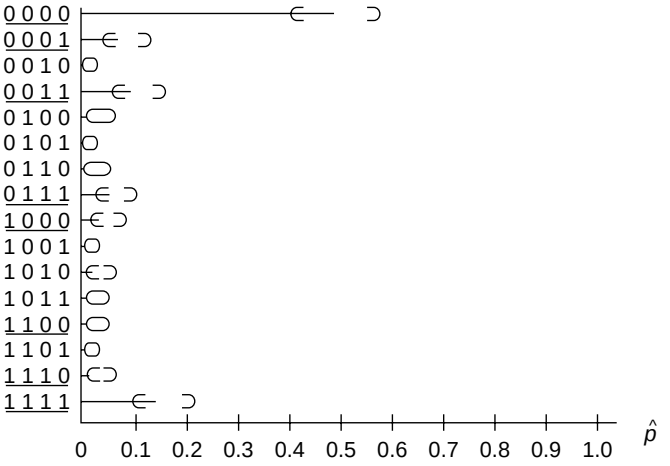


Figure 10.2 Interval estimates of probabilites for situation 2 ($N = 130$)

to 82% in situation 5). In contrast, the fraction of probants showing support for those who are worst-off in situations 1, 2, 3 and 6 is substantially lower and fairly stable between 40 per cent and 50 per cent. Only between 10 per cent and 20 per cent of the respondents do not allocate funds to the worst-off individuals (event 1111). Apart from situation 1, it is this event which carries the second-highest relative frequency.

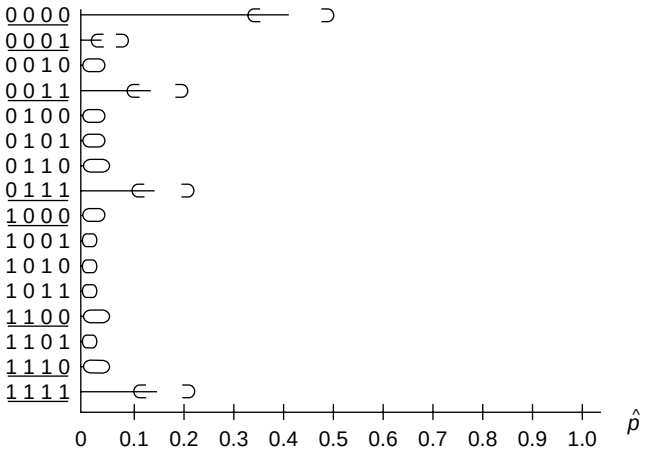


Figure 10.3 Interval estimates of probabilites for situation 3 ($N = 130$)

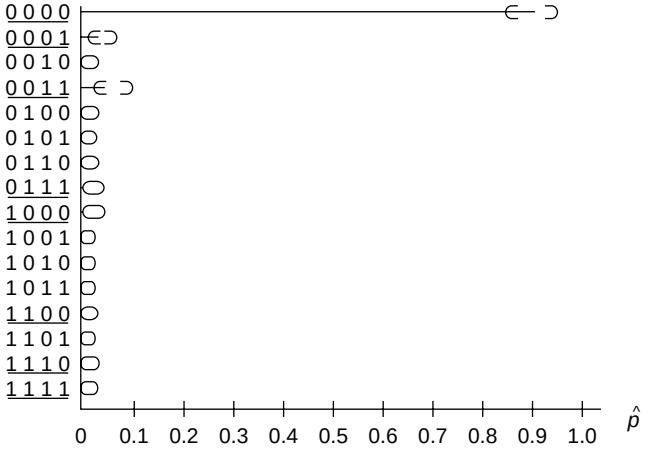


Figure 10.4 Interval estimates of probabilites for situation 4 ($N = 130$)

In situations 1 and 3, approximately 35 per cent of the probants eventually reconsider their support of the worst-off individuals, as the group of those who are better off under the alternatives presented grows. The corresponding magnitudes for situations 2 and 6 are 23 per cent and 28 per cent. A look at the details concerning the distribution of the probability mass over the events 0111, 0011 and 0001 reveals further differences between situations 1 and 3 on the one hand, and 2 and 6 on the other.

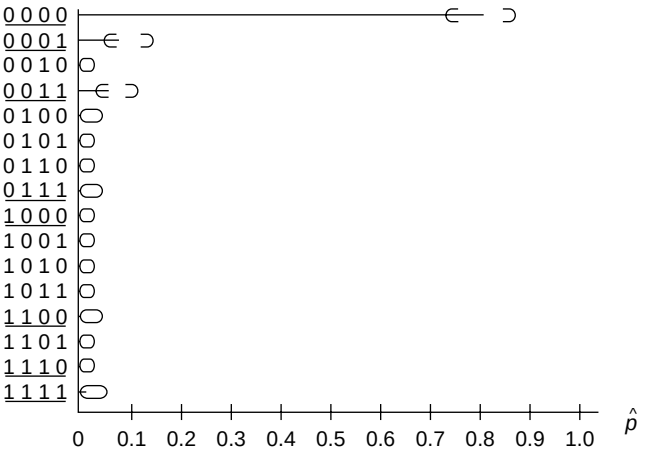


Figure 10.5 Interval estimates of probabilites for situation 5 ($N = 130$)

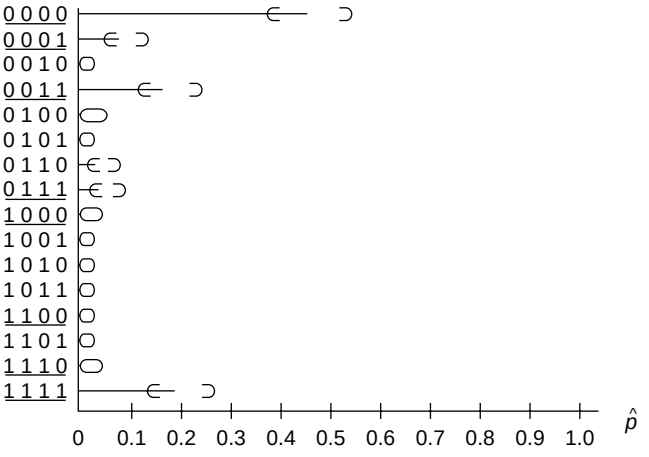


Figure 10.6 Interval estimates of probabilities for situation 6 ($N = 130$)

Figure 10.7 provides a more aggregate presentation of the information in Table 10.2 and Figures 10.1 to 10.6. The estimated probability of deciding in accordance with the Rawlsian equity axiom is measured on the horizontal axis in Figure 10.7. The probability of either initially or eventually deciding in a non-Rawlsian way is measured on the vertical axis. Finally, the estimated probability of not deciding according to these patterns is measured by the vertical or horizontal distance from the point representing a situation to the

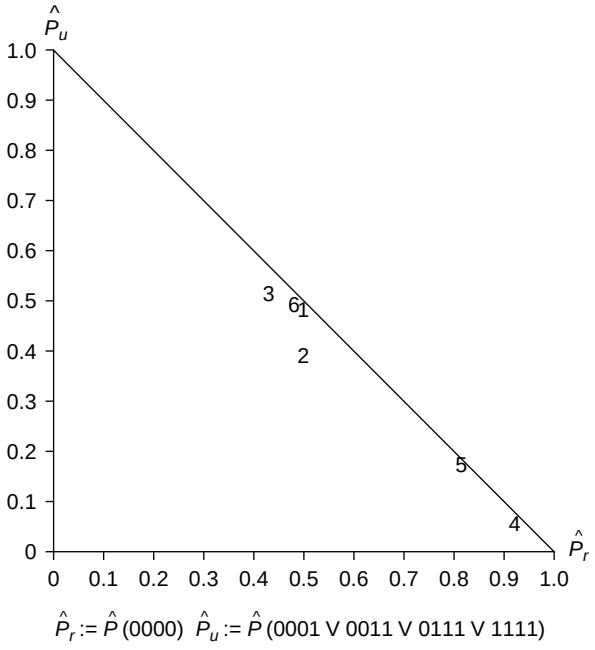


Figure 10.7 Situations 1–6 in $(\hat{P}_r, \hat{P}_u)$-space ($N = 130$)

hypotenuse of the equilateral triangle. This distance reflects the probability of what we have characterized as inconsistent decisions. Notice that our probants were most likely to decide inconsistently in situation 2. Although to a smaller degree, we also observed inconsistent decisions in situation 3. In the other situations very few decisions were inconsistent. We return to a closer examination of inconsistency later in this section.

Two distinct clusters of decision patterns are apparent in Figure 10.7. Cluster A consists of situations 1, 2, 3 and 6, while situations 4 and 5 constitute cluster B. As pointed out in section 3, the situations in cluster B were introduced mainly to check for the logic and consistency of probants' answers. We are primarily interested in the situations in cluster A which appear to have similar decision patterns, at least on the aggregation level of Figure 10.7. Nevertheless, we will carry out a formal test for differences in decision patterns found in cluster A. This is the issue to which we now turn.

Context-dependence of decision pattern

The multi (local) modal nature of the distributions in Figures 10.1, 10.2, 10.3 and 10.6 suggests that testing a hypothesis involving only the location parameters of densities is inadequate. A more comprehensive analysis

revealing subtle differences in the characteristics of the distributions is called for. We test the hypothesis of pairwise equality of multinomial distributions using a Pearson-type χ^2-test. Details concerning the justification of this choice and the description of the test itself can be found in Jungeilges and Theisen (2003).

Let p_{ij} denote the probability of realizing outcome i with $i \in \{0, 1, 2, \ldots, 15\}$ in situation s with $s \in S := \{1, 2, 3, 6\}$. For each of the 6 pairs in the set $\{(s, s') | s, s' \in S \wedge s < s'\}$, we then test the hypotheses:

$$H_0 : p_{is} = p_{is'} \quad \forall i$$

$$H_1 : p_{is} \neq p_{is'} \quad \text{for at least one } i$$

The results for the six χ^2-tests are shown in Table 10.3. In each case, we list the realization of the test statistic along with the associated p-value and the degrees of freedom. Convincing evidence against the null hypothesis of identical densities is found only for the pairs (1,2) and (1,6). By decomposing the χ^2-test statistic into the contributions by each single event, we can identify the importance of events for the rejection of the null hypothesis. Again, for more details on this disaggregation the reader is referred to Jungeilges and Theisen (2003).

For the case of situation 1 versus situation 2, our decomposition of the test statistic indicates that event '7' (corresponding to 0111) accounts for approximately 20 per cent of the value assumed by the test statistic leading to the rejection of H_0. This rejection does not come as a surprise, considering the distance between the points 1 and 2 in Figure 10.7 as well as the difference between the relative frequencies for the event 0111 in Figures 10.1 and 10.2. From the proximity of the points 1 and 6 in Figure 10.7 it is surprising that H_0 is clearly rejected in the test involving situations 1 and 6. The explanation

Table 10.3 Results of χ^2-tests for the equality of two multinomials

	Sit. 2	*Sit. 3*	*Sit. 6*
Sit. 1	24.8583	8.6319	23.5764
	0.0155	0.6558	0.0050
	12	11	9
Sit. 2		18.6239	12.9282
		0.1352	0.2980
		13	11
Sit. 3			16.3258
			0.1295
			11

again lies in the difference between the relative frequencies of the event 0111 (cf. Figures 10.1 and 10.6). This event accounts for nearly 60 per cent of the χ^2 statistic.

In the baseline question, the majority of the respondents reveal themselves as either Rawlsians (40%–50%) or non-Rawlsians (10%–20%). The majority of the remaining probants eventually respond to modifications of the situations by leaving the Rawlsian position. This feature holds across all situations 1, 2, 3 and 6. In addition, four out of six formal tests fail to produce significant evidence against the null hypothesis of identical distributional preferences. Using situation 1 as a reference, we find significant evidence against the null hypothesis for situations 2 and 6. In both cases, the strong evidence can mainly be attributed to the events 0001, 0011 and 0111, which all reflect moves away from the Rawlsian position.

Situation 1 is unique in the sense that probants are confronted with a simple quantitative expansion of the group that could benefit if resources are not allocated to the worst-off individual. As indicated by Figure 10.8, the fraction of the respondents maintaining the Rawlsian position is a decreasing function of the number of highly gifted individuals. Due to the limitations of the current design, we cannot claim that the share of Rawlsians settles at some positive value as the number of gifted individuals approaches infinity.

In contrast to the reference situation, in situation 2 the expansion is measured on an ordinal scale. As in situation 1, the share of probants who maintain the Rawlsian position decreases as the number of those benefiting from environmental measures increases. The nature of the decline is, however, different from the one observed in situation 1. In the formal tests, this showed up as significant differences between the density functions.

In situation 6, the fraction of probants not willing to give up basic rights of workers in favour of higher growth decreases as more groups benefit from higher growth. Again, as indicated by formal tests, the nature of the decline differs from what was observed in situation 1. Note the steep decline when the category *civil servants*, denoted as 'c' in situation 6, is introduced.

Education effects

In our sample, we can distinguish between first and third-year students. This enables us to test for the existence of educational-programme effects. In the economics literature, the related question whether economists and non-economists behave differently in experimental as well as in real-world settings has been studied by a number of researchers. For instance, Marwell and Ames (1981) carried out an experiment of the private provision of public goods, where each of N players received an initial endowment. Each player had to decide privately on the fraction of this endowment that she or he would add to a group fund. A multiple of the group fund was then distributed equally among the N members of society, irrespective of the contribution of each individual. The results of this experiement showed that graduate students

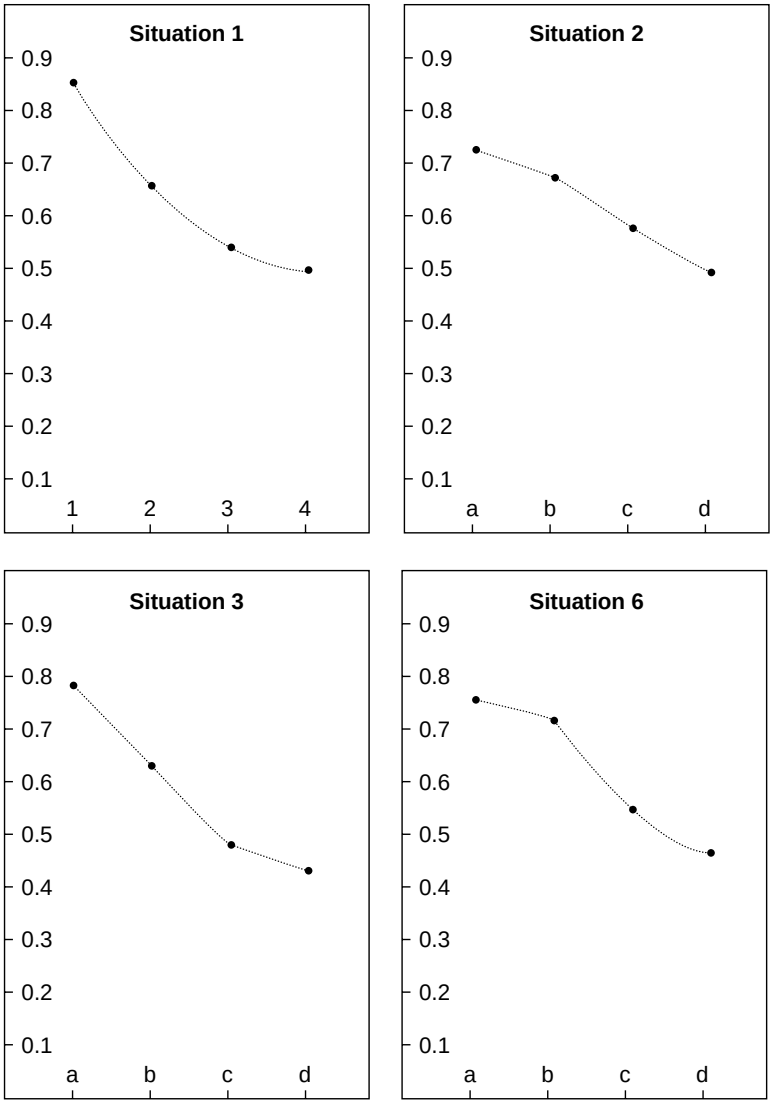


Figure 10.8 Probability of fulfillment of the equity axiom ($N = 130$)

of economics were much more likely to free-ride than students of any other background. Similar results were obtained by Carter and Irons (1991) in an ultimatum bargaining experiment. (This is a game between 2 players. Player 1 has the right to propose a split of a given amount of money. Player 2 has the right to either accept or to reject this offer. If player 2 rejects the offer, then

the pay-off to both players is 0 and the game is over.) In a threshold game of public goods provision, Cadsby and Maynes (1998) found results pointing in the same direction. Similar results have been reported by Frank, Gilovich and Regan (1993a) in a free-rider experiment, and by Frank and Schultze (2000) in a study of participants' honesty in an experimental principal-agent context. Selten and Ockenfels (1998) found, in an experimental solidarity game, that male economics students were less generous than other male students, while there was no significant difference among female students. Contrasting results were obtained by Yetzer, Goldfarb and Poppen (1996), who in a 'lost-letter' experiment found that economics students were more honest than other students. Also Laband and Beil (1999), who studied 'real-world' decisions, concluded that professional economists were no more selfish than non-economic professionals.

In several of the papers mentioned above, a distinction is made between the *self-selection effects* and the *learning effects* of education. These effects may also play a role in our case. If students with certain interests and attitudes are more likely to apply for admission to the advanced business administration programme (ABA), there will be a self-selection effect. If students' behaviour is modified as a result of being exposed to economic reasoning, there will be an educational effect. In addition, as mentioned in section 4, students who apply to the ABA programme are screened on the basis of their grades earned in the BBA programme. Hence, there may be a *screening effect*. The self-selection effect, the learning effect and the screening effect add up to the *educational-programme-effect*, which, if it is discernible, will manifest itself in different responses from first-year students and third-year students.

It may be difficult to disentangle empirically the three components of the educational effect. The previous literature tends to emphasize the selection effect. In our case, it seems reasonable to expect that the self-selection, screening and learning effects will lead to the result that third-year students possess more distinct abilities and skills needed to solve economic decision problems such as those in our experiment. Furthermore, the more experienced students may also understand better the rationale for carrying out the experiment. Consequently, we formulate the following hypotheses:

1 Third-year students are more likely to participate in the experiment than are first-year students.
2 Third-year students will have a higher propensity than first-year students to comply with the experimenter.
3 Third-year students will be more likely than first-year students to make consistent decisions.

These hypotheses are now tested with the data. Let p_i denote the probability for the event 'a student in the *i*-th year participates in the experiment',

$i = 1, 3$. Testing the first hypothesis just listed amounts to carrying out the following one-sided test:

$$H_0 : p_1 \geq p_3$$
$$H_1 : p_1 < p_3$$

A test of these hypotheses is based on observations taken on N probants. Each subject can be related to one of two classes and one of two populations. After categorization, the absolute frequencies f_{ij} are presented in 2×2 contigency tables of the format:

	Class 1	***Class 2***	$\sum$
Population 1	f_{11}	f_{12}	n_1
Population 2	f_{21}	f_{22}	n_2
$\sum$	c_1	c_2	$N = n_1 + n_2$

Under the assumptions that one faces two mutually independent random samples, one-sided hypotheses concerning p_1 and p_3 are tested using the statistic:

$$T_1 = \frac{\sqrt{N}(f_{11}f_{22} - f_{12}f_{21})}{\sqrt{n_1 n_2 c_1 c_2}}$$

which follows a standard normal distribution asymptotically.

Data used to test the hypothesis that third-year students are more likely to participate in the experiment are organized in Table 10.4. We obtain the estimated participation probabilities $\hat{p}_1 = 0.78$ and $\hat{p}_3 = 0.91$. The realization of the test statistic $\hat{T}_1 = -2.3216$ implies that, at a significance level of $\alpha = 0.05$, we reject the null hypothesis. A comparison of $\hat{T}_1$ with the critical value for the one-sided test $x_{1-\alpha} = -1.6452$ shows that the result is indeed highly significant. Consequently, we find convincing evidence that third-year students show a higher propensity to participate in the experiment.

Let us now consider the 130 students who participated in the experiment and test for differences in compliance between first and third-year students

Table 10.4 Testing for differences in participation probabilities

	Participation	*Withdrawal*	*Total*
Population 1 (1st year)	66	19	85
Population 2 (3rd year)	64	6	70
Total	130	25	155

Table 10.5 Testing for differences in compliance probabilities

	Compliance	*Non-compliance*	*Total*
Population 1 (1st year)	49	17	66
Population 2 (3rd year)	61	3	64
Total	110	20	130

(cf. hypothesis 2). By definition, a probant complies with the experiment if (s)he returns an answer sheet providing complete data on all demographic items. If we let p_i denote the probability for the event 'a student in the i-th year complies with experiment', then we can formalize the hypothesis (2) as:

$$H_0 : p_1 \geq p_3$$
$$H_1 : p_1 < p_3$$

Under the assumptions stated above, this one-sided hypothesis can again be tested using the test statistic T_1 based on the absolute frequencies in Table 10.5. The estimated compliance probability for the third-year students $\hat{p}_3 = 0.95$ exceeds the probability for the first year $\hat{p}_1 = 0.74$. The test statistic $\hat{T}_1$ equals -3.3288. We clearly reject H_0 at all conventional significance levels. In other words, our data clearly support the conjecture that third-year students show a higher propensity to comply with the experiment.

Next, let us turn to the third hypothesis concerning students' ability to decide consistently, in the sense defined at the beginning of section 5. The probability for the event 'a student in the i-th year provides a response being an element of C (cf. Definition 1) in at least one of the six situations' is denoted by p_i. Therefore, we test the hypotheses

$$H_0 : p_1 \leq p_3$$
$$H_1 : p_1 > p_3$$

Based on the information in Table 10.6, we obtain the estimates $\hat{p}_1 = 0.21$ and $\hat{p}_3 = 0.27$ suggesting that third-year students are more likely to decide inconsistently. However, a realization of $\hat{T}_1 = -0.7157$ does not justify the rejection of the null hypothesis. We carried out the same test for the subsample of the 110 probants who provided complete demographic information. The resulting value on the test statistic, -0.7143, is amazingly similar. This indicates that those who did not provide complete answers were as consistent as those who did provide the full demographic information.

Several of the papers referred to at the beginning of this subsection suggest that students who for a long time have been exposed to the study of competitive markets may adopt the competitive model as a norm for how a society

Table 10.6 Testing for differences in inconsistent responses

	Inconsistent	*Consistent*	*Total*
Population 1 (1st year)	14	52	66
Population 2 (3rd year)	17	47	64
Total	32	99	130

should allocate resources. Moreover, the self-selection and screening effects may lead to the result that students who exhibit a higher degree of acceptance for the operation of competitive markets are more likely to enter the advanced business administration programme.

Based on such arguments, one may formulate the hypothesis that third-year students are likely to act less in accordance with the Rawlsian equity axiom than are first-year students. Below, we will also test this hypothesis. Notice that the theoretical basis for such a hypothesis appears to be weak. Nevertheless, it seems worthwhile to test for differences in distributional judgements held by first and third-year students.

An informal assessment of Tables 10.7 and 10.8 indeed suggests that first-year students tend to be more Rawlsian. Note, in particular, differences for the events '0' (0000) and '15' (1111). Again, to arrive at a firmer conclusion, we have to subject our conjecture to a formal test. For this purpose, let $p_{is}^{1st}(p_{is}^{3rd})$ denote the probability that a first-year student chooses according to event i with $i \in \{0, 1, 2, \ldots, 15\}$ in situation s with $s \in S^f := \{1, 2, 3, 4, 5, 6\}$. For each $s \in S^f$, we then test the hypotheses:

$$H_0 : p_{is}^{1st} = p_{is}^{3rd} \quad \forall i$$
$$H_1 : p_{is}^{1st} \neq p_{is}^{3rd} \quad \text{for at least one } i$$

The results for the χ^2-tests in Table 10.9 provide very little evidence against the null hypothesis. In other words, we do not find significant differences between the density functions for the two sub-populations (first-year and third-year students). This finding is in line with the conclusion of Jacobsen (2001), who in a study of the political attitudes of students found that third-year students held the same attitudes as first-year students. Jacobsen's results are based on data collected at the same university college at which we carried out our experiment.

Providers and non-providers of demographic information

The investigation in the previous subsection was based on a sample of $N = 130$ responses. However, as mentioned in section 4, only 110 out of the 130 valid records contained complete information about demographic characteristics

Table 10.7 Estimated densities for first-year respondents ($N = 49$)

Cases	*Decimal*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0.000	0.531	0.551	0.531	0.939	0.898	0.531
0 0 0 1	1.000	0.041	0.061	0.082	0.000	0.061	0.102
0 0 1 0	2.000	0.000	0.000	0.000	0.000	0.000	0.000
0 0 1 1	3.000	0.163	0.102	0.163	0.061	0.041	0.143
0 1 0 0	4.000	0.000	0.041	0.020	0.000	0.000	0.000
0 1 0 1	5.000	0.000	0.000	0.000	0.000	0.000	0.000
0 1 1 0	6.000	0.020	0.000	0.020	0.000	0.000	0.000
0 1 1 1	7.000	0.163	0.020	0.082	0.000	0.000	0.061
1 0 0 0	8.000	0.000	0.061	0.020	0.000	0.000	0.020
1 0 0 1	9.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10.000	0.000	0.020	0.000	0.000	0.000	0.000
1 0 1 1	11.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 0 0	12.000	0.000	0.020	0.000	0.000	0.000	0.000
1 1 0 1	13.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 1	15.000	0.082	0.122	0.082	0.000	0.000	0.143

Table 10.8 Estimated densities for third-year respondents ($N = 61$)

Cases	*Decimal*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0.000	0.443	0.459	0.328	0.951	0.803	0.426
0 0 0 1	1.000	0.049	0.082	0.033	0.016	0.082	0.049
0 0 1 0	2.000	0.016	0.000	0.016	0.000	0.000	0.000
0 0 1 1	3.000	0.098	0.098	0.148	0.033	0.066	0.197
0 1 0 0	4.000	0.000	0.000	0.000	0.000	0.016	0.033
0 1 0 1	5.000	0.000	0.000	0.016	0.000	0.000	0.000
0 1 1 0	6.000	0.000	0.033	0.016	0.000	0.000	0.049
0 1 1 1	7.000	0.230	0.082	0.213	0.000	0.016	0.016
1 0 0 0	8.000	0.033	0.033	0.000	0.000	0.000	0.000
1 0 0 1	9.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 1	11.000	0.000	0.016	0.000	0.000	0.000	0.000
1 1 0 0	12.000	0.000	0.000	0.033	0.000	0.000	0.000
1 1 0 1	13.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14.000	0.000	0.016	0.016	0.000	0.000	0.000
1 1 1 1	15.000	0.131	0.180	0.180	0.000	0.016	0.230

and other background variables. In relating decisions to characteristics, we have to rely on the subset of only 110 complete observations. To relate results of such analyses to the densities estimated on the basis of the full sample, we have to establish that probants who refused to provide demographic or

Table 10.9 Results of χ^2-tests for the equality of two multinomials

	Sit. 1	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
Statistic	6.23944	11.77742	15.41382	1.29089	3.16405	9.95855
p-value	0.51209	0.38060	0.16432	0.52443	0.67471	0.19093
df	7	11	11	2	5	7

Table 10.10 Estimated densities for first-year students complete records ($N = 49$)

Cases	*Decimal*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0.000	0.531	0.551	0.531	0.939	0.898	0.531
0 0 0 1	1.000	0.041	0.061	0.082	0.000	0.061	0.102
0 0 1 0	2.000	0.000	0.000	0.000	0.000	0.000	0.000
0 0 1 1	3.000	0.163	0.102	0.163	0.061	0.041	0.143
0 1 0 0	4.000	0.000	0.041	0.020	0.000	0.000	0.000
0 1 0 1	5.000	0.000	0.000	0.000	0.000	0.000	0.000
0 1 1 0	6.000	0.020	0.000	0.020	0.000	0.000	0.000
0 1 1 1	7.000	0.163	0.020	0.082	0.000	0.000	0.061
1 0 0 0	8.000	0.000	0.061	0.020	0.000	0.000	0.020
1 0 0 1	9.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10.000	0.000	0.020	0.000	0.000	0.000	0.000
1 0 1 1	11.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 0 0	12.000	0.000	0.020	0.000	0.000	0.000	0.000
1 1 0 1	13.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 1	15.000	0.082	0.122	0.082	0.000	0.000	0.143

other background information did in fact exhibit the same decision pattern as those who provided this information.

Tables 10.10 and 10.11 allow for an informal comparison of the decision pattern prominent in the two groups of respondents. Since only three third-year students did not provide full information, we confine the comparison to the first-year students. Considering the width of the confidence intervals in Figures 10.1 to 10.6 (based on $N = 130$ observations) and the small sample sizes in Tables 10.9 and 10.10, it seems unlikely that the hypothesis concerning the equality of the densities associated with the two groups could be rejected in any of the situations.

Nevertheless, let us carry out a formal test of the hypothesis just addressed. For this purpose, let $p_{is}^{d}(p_{is}^{\neg d})$ denote the probability that a first-year student who provided complete (no) information chooses according to event i with $i \in \{0, 1, 2, \ldots, 15\}$ in situation s with $s \in S^f := \{1, 2, 3, 4, 5, 6\}$. For each $s \in S^f$,

Table 10.11 Estimated densities for first-year students with incomplete records ($N = 17$)

Cases	*Decimal*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0.000	0.647	0.529	0.529	0.706	0.647	0.471
0 0 0 1	1.000	0.000	0.118	0.000	0.118	0.176	0.118
0 0 1 0	2.000	0.000	0.000	0.000	0.000	0.000	0.000
0 0 1 1	3.000	0.118	0.118	0.118	0.059	0.059	0.118
0 1 0 0	4.000	0.000	0.000	0.000	0.000	0.000	0.000
0 1 0 1	5.000	0.000	0.000	0.000	0.000	0.000	0.000
0 1 1 0	6.000	0.059	0.000	0.000	0.000	0.000	0.059
0 1 1 1	7.000	0.059	0.059	0.118	0.059	0.000	0.059
1 0 0 0	8.000	0.000	0.000	0.000	0.059	0.000	0.000
1 0 0 1	9.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10.000	0.000	0.059	0.000	0.000	0.000	0.000
1 0 1 1	11.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 0 0	12.000	0.000	0.000	0.000	0.000	0.059	0.000
1 1 0 1	13.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14.000	0.000	0.118	0.059	0.000	0.000	0.000
1 1 1 1	15.000	0.118	0.000	0.176	0.000	0.059	0.176

Table 10.12 Results of χ^2-tests for the equality of two multinomials

	Sit. 1	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
Statistic	2.9768	11.7274	6.7347	12.3096	8.6521	3.5007
p-value	0.7036	0.2291	0.5655	**0.0152**	**0.0704**	0.7439
df	5	8	8	4	4	6

we then test the hypotheses:

$$H_0 : p_{is}^{d} = p_{is}^{\neg d} \quad \forall i$$

$$H_1 : p_{is}^{d} \neq p_{is}^{\neg d} \quad \text{for at least one } i$$

Apart from situation 4 – and 5 which is a borderline case – Table 10.12 does not give convincing evidence against the hypothesis that the first-year student who did not provide the complete information exhibited the same decision behaviour as those first-year students who revealed their demographic and background characteristics. The analysis of the extended output for the test reveals the source of the relatively large realization of the test statistic in the case of situation 4: two of the 17 probants who did not reveal full information decided to choose according to 0001, that is they eventually catered to the wine drinkers.

Gender effects

Despite early experimental evidence pointing in the direction of existing gender effects in coordination games (for example Rapoport and Chammah, 1965), gender as a factor influencing behaviour has largely been ignored in economic theory. On the contrary, gender is viewed as a factor determining behaviour in psychology. According to a prominent view among psychologists, females and males perceive the same environment in a very different way (Tannen, 2001; Gray, 1991). Females tend to interpret their surroundings in terms of a network of social relations, perceiving themselves as a part of that network. To contribute to the welfare of the entire network is viewed as a moral obligation. Gilligan (1982) and Rosener (1990) point out that males, on the other hand, interpret their environments in terms of hierarchical relationships and tend to focus more on the rights of the individual. On the basis of such views, it is reasonable to hypothesize that females and males behave differently when making choices in economic contexts, with females being more concerned about fairness and the welfare of their reference group.

A review of work on gender effects in experimental economics shows that existing evidence is not conclusive. Results vary according to the type of experiment, the type of design and the composition of the sample. In *prisoner's dilemma*-type games, females tend to be more cooperative than males, at least in the initial rounds of the games. In Frank, Gilovich and Regan (1993b), the difference in probabilities for defection between males and females equals 0.24. According to Ortmann and Tichy (1999), such differences vanish as experience is accumulated. Similar evidence is presented for market experiments by Mason and Philips (1991). In addition, they find less stable decision patterns for females than for males, at least in the initial phase of their games. Very different findings have been published in the context of *dictator games*. (A game played between two players. Player 1 has the right to propose a split of a given amount of money between him and player 2. Player 2 has to accept the split.) While, for instance, Bolton and Katok (1995) fail to establish significant gender effects, data from an experiment relying on a different design by Eckel and Grossman (1998) allow for the identification of a gender effect. Based on data from a modified dictator game, Andreoni and Vesterlund (2001) arrive at the conclusion that gender effects are complex: Men are more likely to be perfectly selfish or perfectly altruistic, while women tend to act more evenly. Findings in the context of *ultimatum games* indicate that players expect females to be more cooperative than males. Moreover, offers made by females tend to be rejected more often than offers made by males (Solnick, 2001; Eckel and Grossman, 1998).

Mixed evidence is also found in the context of *public goods experiments*. For example, the finding of Brown-Kruse and Hummels (1993) that groups composed of only females tend to contribute less than all-male groups is reversed in results reported by Novell and Tinkler (1994). Under an experimental

setting which excludes the possibility for probants to influence or being influenced by group members, Sell, Griffith and Wilson (1993) find that gender is not activated at all. Allowing for this possibility of influence by a change in experimental conditions, Sell (1997) finds that women were not more cooperative when placed in all-female groups than in groups with a mixed gender composition. On the other hand, for men who were put in mixed gender groups she found a higher propensity to cooperate than for men in all-men groups. It has been demonstrated by Cadsby and Maynes (1998) that a change in the composition of the sample from a selection of undergraduate students to a true random sample renders the results insignificant. In the context of a *solidarity game*, Selten and Ockenfels (1998) establish gender as a significant behavioural determinant. Females tend to contribute more than males. In an experiment designed to highlight possible determinants of corruptability, Frank and Schultze (2000) fail to find a gender effect for pooled data, but obtain evidence for the existence of such an effect once they focus on subgroups of the student population. Contrasting the dominance of self-interest for economists versus non-economists they find stronger evidence for choice behaviour driven by self-interest for males than for females. Finally, in a questionnaire based study of a standard axiom of distributional comparison, Amiel and Cowell (2000) find significant gender effects, especially on items involving aspects of risk. Their male probants tend to interpret equalizing transfers as inequality-reducing and risk-reducing more often than female probants. The responses of men also are more consistent with the Principle of Transfers. The signifiance of gender has, of course, been considered outside the area of experimental economics. For an interesting discussion in the context of development and environmental economics, see Agarwal (2000).

Turning to the evidence from our survey experiment, a first comparison of densities for females and males in Tables 10.13 and 10.14 – situation by situation – suggests that the incidence of Rawlsian type choices seem to be more prominent among females than among males. To shed light on the validity of this conclusion, a formal test of the hypothesis is necessary.

Let $p^f_{is}(p^m_{is})$ denote the probability that a female (male) student responds according to event i with $i \in \{0, 1, 2, \ldots, 15\}$ in situation s with $s \in S^f := \{1, 2, 3, 4, 5, 6\}$. For each $s \in S^f$ we then test the hypotheses:

$$H_0 : p^f_{is} = p^m_{is} \quad \forall i$$

$$H_1 : p^f_{is} \neq p^m_{is} \quad \text{for at least one } i$$

The results for the six χ^2-tests are given in Table 10.15. The lowest p-value equals 0.13, indicating that the observed differences between females and males are not significant.

Table 10.13 Estimated densities for female respondents ($N = 56$)

Cases	*Decimal*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0.000	0.536	0.661	0.536	0.982	0.875	0.536
0 0 0 1	1.000	0.054	0.054	0.054	0.000	0.036	0.107
0 0 1 0	2.000	0.000	0.000	0.000	0.000	0.000	0.000
0 0 1 1	3.000	0.179	0.089	0.143	0.018	0.036	0.161
0 1 0 0	4.000	0.000	0.018	0.018	0.000	0.018	0.018
0 1 0 1	5.000	0.000	0.000	0.018	0.000	0.000	0.000
0 1 1 0	6.000	0.018	0.000	0.018	0.000	0.000	0.036
0 1 1 1	7.000	0.143	0.054	0.089	0.000	0.018	0.018
1 0 0 0	8.000	0.000	0.036	0.000	0.000	0.000	0.000
1 0 0 1	9.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 1	11.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 0 0	12.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 0 1	13.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14.000	0.000	0.000	0.018	0.000	0.000	0.000
1 1 1 1	15.000	0.071	0.089	0.107	0.000	0.018	0.125

Table 10.14 Estimated densities for male respondents ($N = 54$)

Cases	*Decimal*	*Sit. 1*	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
0 0 0 0	0.000	0.426	0.333	0.296	0.907	0.815	0.407
0 0 0 1	1.000	0.037	0.093	0.056	0.019	0.111	0.037
0 0 1 0	2.000	0.019	0.000	0.019	0.000	0.000	0.000
0 0 1 1	3.000	0.074	0.111	0.167	0.074	0.074	0.185
0 1 0 0	4.000	0.000	0.019	0.000	0.000	0.000	0.019
0 1 0 1	5.000	0.000	0.000	0.000	0.000	0.000	0.000
0 1 1 0	6.000	0.000	0.037	0.019	0.000	0.000	0.019
0 1 1 1	7.000	0.259	0.056	0.222	0.000	0.000	0.056
1 0 0 0	8.000	0.037	0.056	0.019	0.000	0.000	0.019
1 0 0 1	9.000	0.000	0.000	0.000	0.000	0.000	0.000
1 0 1 0	10.000	0.000	0.019	0.000	0.000	0.000	0.000
1 0 1 1	11.000	0.000	0.019	0.000	0.000	0.000	0.000
1 1 0 0	12.000	0.000	0.019	0.037	0.000	0.000	0.000
1 1 0 1	13.000	0.000	0.000	0.000	0.000	0.000	0.000
1 1 1 0	14.000	0.000	0.019	0.000	0.000	0.000	0.000
1 1 1 1	15.000	0.148	0.222	0.167	0.000	0.000	0.259

Table 10.15 Results of χ^2-tests for the equality of two multinomials

	Sit. 1	*Sit. 2*	*Sit. 3*	*Sit. 4*	*Sit. 5*	*Sit. 6*
Statistic	10.63281	16.20589	14.77057	3.11082	5.90107	7.91632
p-value	0.15546	0.13366	0.19324	0.21110	0.31596	0.34003
df	7	11	11	2	5	7

6 Binary response models

The hypotheses tested in section 5 involved the entire density of the discrete random variable D under different situations and/or for different subpopulations of probants. We now turn to model two specific aspects of the decisions taken by the respondents: (i) the fulfillment of the equity axiom and (ii) the reconsideration of an initial Rawlsian position. The probabilities for the events (i) and (ii) are modelled as functions of demographic characteristics and other variables reflecting the background of the respondents.

Modelling the propensity to fulfill the *equity axiom*

Econometric model

Let us consider the decision process of a probant facing one of the situations in the experimental setup. An elementary outcome of the experiment takes the form of a binary string consisting of four elements, each element indicating the decision made in response to one of the four questions in a given situation. Defining:

$$d_{(\bullet,j)} = \begin{cases} 0 & \text{if respondent allocates funds to the disadvantaged} \\ & \text{in question } j \\ 1 & \text{if otherwise} \end{cases}$$

with $j \in \{1, \ldots, 4\}$, we can express the ith elementary outcome of the experiment as:

$$\omega_i = (d_{(i,1)}, d_{(i,2)}, d_{(i,3)}, d_{(i,4)})$$

The element $\omega_i = (0000)$, for instance, indicates that a respondent allocated the funds to the disadvantaged individual(s) in the baseline question. As more individuals were added to the group of those who are better off under all circumstances, the respondent does not reconsider this decision. Let Ω denote the set of all possible outcomes:

$$\Omega = \{\omega_0, \ldots, \omega_i, \ldots, \omega_{15}\}$$

In what follows, we assume the existence of the probability space $(\Omega, \sum, P)$, where $\sum$ denotes the σ-algebra of events containing all events to which we can assign probabilities. P indicates the probability measure. In the following, we shall also rely on the decimal representation of the elementary event ω_i.

Definition 2 $E = \{\omega_i \mid d_{i1} = 0 \wedge i \in \{0, \ldots, 15\}\}$

The set E contains all elementary events in which the *equity axiom* is fulfilled in the baseline question of a given situation where the fund is allocated to the individuals who are worst off. That is, E contains the eight elementary events associated with the decimal representation '0' to '7' in Table 10.2. In the next step, we define a binary random variable Y.

Definition 3 Let $I_E(\bullet)$ denote the indicator function:

$$I_E(x) = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{if } x \notin E \end{cases}$$

Then

$$Y(\omega_i) = I_E(\omega_i), \quad \omega_i \in \Omega$$

It is assumed that the discrete random variable Y follows a *Bernoulli* distribution.

Assumption 1 $Y \sim B(p) = p^y(1-p)^{1-y} I_{\{0,1\}}(y)$ where $p = P(Y = 1) = P(E)$ and y denotes a realization of the random variable Y.

Next, suppose that the probability of p for the event E depends on a set of covariates which constitute the vector $z \in \mathbb{R}^c$. In the context at hand, the covariates are characteristics of the respondent.

Assumption 2 Let $F: \mathbb{R} \to [0, 1]$ be a monotonically increasing function and $\theta \in \mathbb{R}^c$. Then $p = F(\theta' z)$.

As a consequence, the binary random variable Y follows the Bernoulli density:

$$Y \sim B(F(\theta' z)) = F(\theta' z)^y (1 - F(\theta' z))^{1-y} I_{\{0,1\}}(y)$$

Let Y_k reflect the decision of the k-th respondent faced with one of the six situations:

$$Y_k = \begin{cases} 1 & \text{equity axiom fulfilled} \\ 0 & \text{otherwise} \end{cases}$$

and z_k denotes the characteristics observed for respondent k for $k = (1, \ldots, N)$. Assuming that the elements of the sample $(Y_1, \ldots, Y_N)$ are *independently* and *identically* Bernoulli distributed random variables, the log-likelihood function takes the form:

$$\ln\ L(\theta \mid y_1, \ldots, y_N) = \sum_{k=1}^{N} y_k\ \ln\ F(\theta' z_k) + (1 - y_k) \ln(1 - F(\theta' z_k)) \tag{1}$$

The ML estimator

$$\hat{\theta} = \arg\ \max(\ln\ L(\theta \mid y_1, \ldots, y_N)) \in \Theta$$

exists under mild regularity conditions. It is well-established for certain specifications of F that the numerical determination of the ML estimator is feasible. This holds, for example, for the logistic distribution $L(0, 1)$:

$$F(\theta' z_k) = [1 + e^{-\theta' z_k}]^{-1}$$

Under this specification, the Hessian is guaranteed to be negative definite. The goal function (1) is therefore globally concave. A Newton–Raphson-type numerical procedure will be employed to generate the ML estimates for the logit models. In estimating the variance $V[\hat{\theta}]$, we followed standard strategies outlined, for instance, in Green (1990) or Davidson and MacKinnon (1993).

Covariates of the logit models

The elements of the covariate vector z that will be used in the estimation of logit equations are defined in Table 10.16. In addition to the gender variable

Table 10.16 Definition of variables used in the logit analysis

Acronym	*Description*
GROUP	Probant is a (1=) first-year (2=) third-year student
AGE	Age of probant measured in years
SEX	1 = probant is female, 2 = probant is male
PAR	Probant classifies parents as 1 = workers, 2 = craftsmen, 3 = qualified workers, 4 = employees in public administration, 5 = working in the private sector, 6 = self-employed
MAJOR	Probant's major subject of study classified as
-+	1 = business administration, 2 = mathematics, 3 = economics, 4 = other
JOEX	Probant has 1 = job experience, 2 = no job experience
PPLI	Probant expects that 1 = 5%, 2 = 25%, 3 = 50%, 4 = 75%, 5 = 95% of the citizens will earn less than he or she does in 2010

SEX and the educational variable GROUP already considered in section 5, we introduce four additional explanatory variables: PPLI, PAR, AGE and JOEX.

First consider the rationale for including the PPLI variable which measures the respondents expected position in a future income distribution. If people are Rawlsians, choosing a social welfare function behind the veil of ignorance, there is certainly no argument for introducing this explanatory variable. Gaertner, Jungeilges and Neck (2001) argued however that 'individuals in experiments do not put themselves under a "veil of ignorance", but consciously take into account their personal interests when making judgements'. By including the PPLI variable we allow for testing this presumption.

The variable PAR reflects parental background. Since we want to obtain a parsimonious measure of parental background, probants were asked about their *perceived* parental background. Using a self-perceived measure of socio-economic background we avoid the complexities involved in aggregating the professional status of the two parents to a single variable. This variable is meant to capture the possibility that values are passed on from the older to the younger generation. It may also capture the effect of solidarity within the family. We suspect that a probant coming from a family of workers will share the values of workers and be more likely to adhere to the Rawlsian equity principle than one coming from another social background. We include a job experience variable, JOEX, mainly in order to test whether job experience matters in choice contexts where workers' rights are at stake. Probants with work experience are expected to emphasize workers' rights more than those with no working experience. As individuals grow older, their attitudes may change. The AGE variable is included to capture this effect. Even if the age variation in our sample is limited, the variable could still have an important impact. Our basis for formulating a more specific hypothesis concerning the effect of AGE is weak. Notice, also, that it is necessary to include an age variable in the model to avoid that the GROUP variable and the job experience variable capture age effects.

Table 10.17 gives the correlation matrix for the explanatory variables that will be used in the logit analysis. Age and job experience are the two variables

Table 10.17 Correlation matrix for explanatory variables

	GROUP	*SEX*	*AGE*	*PAR*	*MAJOR**	*JOEX*	*PPLI*
GROUP	1.0000	0.0020	0.2159	0.0868	0.0858	0.0830	**0.3278**
SEX	0.0020	1.0000	0.1550	0.0968	−0.0941	−0.1650	0.1082
AGE	0.2159	0.1550	1.0000	0.0622	−0.0157	**−0.4172**	0.0752
PAR	0.0868	0.0968	0.0622	1.0000	−0.0079	−0.0036	0.2771
MAJOR	0.0858	−0.0941	−0.0157	−0.0079	1.0000	0.0724	0.0886
JOEX	0.0830	−0.1650	−0.4172	−0.0036	0.0724	1.0000	0.0283
PPLI	0.3278	0.1082	0.0752	0.2771	0.0886	0.0283	1.0000

*Due to insufficient variation, the variable MAJOR is not used in the subsequent analysis.

displaying the highest correlation. Notice also the relatively high positive correlation between PPLI and GROUP. This is likely to be due to the fact that third-year students are a more strongly selected group than first-year students. That is, the students in the two-year study programme who obtain low grades are not admitted to the third year. The better students who are admitted to the third year are likely to have better income prospects. Finally, notice that the correlation between expected position in the income distribution and parental background is also relatively high.

Model selection strategy

Due to the fact that there is no theory available to guide us in the specification of the logit models, we chose an *all-possible-model* selection strategy. Apart from providing a parsimonious statistical model for $P(E)$, this also has the potential to give clues towards factors that might play a crucial role in theoretical explanations. Given the set of covariates $z = (GROUP, SEX, AGE, PAR, JOEX, PPLI)$ one can specify and estimate $2^6 - 1 = 63$ different models, with the intercept being forced into each model. After estimating all l variable models, $l = 1, \ldots, 6$, they were ranked according to the likelihood-ratio index:

$$LRI(\theta_l) = 1 - \frac{\ln L(\theta_l \mid y)}{\ln L(\theta_0 \mid y)}$$

where θ_l refers to a model containing l covariates and θ_0 denotes the model containing the constant term only. The realizations of the criterion are restricted to the unit interval, with $LRI = 0$ indicating that all slope coefficients $\partial E[Y|z_i]/\partial z_i$ are equal to zero. For details on the nature of this criterion, see Green (1990: 891). By determining for each group of l-variate models the one which – in a likelihood sense – constitutes the strongest improvement over and above the model including the intercept only – say θ_l^* – we generate the increasing sequence:

$$LRI(\theta_1^*) < LRI(\theta_2^*) < LRI(\theta_3^*) < \cdots < LRI(\theta_6^*)$$

On the basis of this sequence, we choose as the final model θ^{**} the one with the smallest dimension fulfilling the requirement:

$$LRI(\theta_{i+1}^*) - LRI(\theta_i^*) < \epsilon$$

with $\epsilon > 0$ and small.

Applying this strategy, the set of all covariates is partioned into a set containing potentially useful explanatory variables and its complement constituted by those variables which were not found to contribute significantly to the model. To substantiate our selection results for a given situation we

proceed as follows. After estimating a full model, that is a model containing all covariates, we tested the hypothesis that the coefficients of those variables which were not contributing significantly according to the selection routine, are simultaneously equal to zero using the Wald test and the Likelihood Ratio test (LR).

Estimation results

For each situation, we report the estimation results for the final logit model θ^{**}. For reasons outlined above, we restrict ourselves to reporting the findings for situations 1, 2, 3 and 6. The ML estimates $\hat{\theta}^{**}$ augmented by the associated standard errors, t-ratios and slope coefficients are reported in Tables 10.18 to 10.21.

For situation 1, the smallest model which could not be improved significantly by the inclusion of additional variables contains four variables. In the presence of the covariates GROUP, AGE and PPLI, the variable which comes closest to having a significant partial effect is SEX. The negative slope coefficient implies that females are more likely to allocate the funds to the handicapped individual than male probants. The model including SEX alone constitutes the best single-variate explanation of the data. The inclusion of GROUP, AGE and PPLI increases the t-ratio of SEX only marginally, but the likelihood of the sample increases considerably. The partial effects of the remaining variables suggest that probants who had made it to the third year tended to decide in accordance with the *equity axiom* less often than students who just started their education. *Ceteris paribus*, probants who hold high expectations with respect to their position in a future income distribution are less likely to allocate the funds to the handicapped child. In addition, the estimation results hint at a positive age effect. The evidence in support of the education effect, the income expectation effect and the age effect is weak according to conventional standards.

Table 10.18 ML estimates of logit model for $P(E)$ in situation 1

Variable	*Estimate*	*Std. Dev.*	*t-ratio*	*Slope*
CONST	1.93109	3.93456	0.49080	0.16999
GROUP	−0.94839	0.79494	−1.19303	−0.08349
SEX	−1.15666	0.66663	**−1.73509**	−0.10182
AGE	0.22514	0.20084	1.12098	0.01982
PPLI	−0.47154	0.44112	−1.06896	−0.04151

Log-Likelihood: −37.91059 *LRI*: 0.096

Table 10.19 ML estimates of logit model for $P(E)$ in situation 2

Variable	*Estimate*	*Std. Dev.*	*t-ratio*	*Slope*
CONST	7.00961	1.72269	4.06900	1.06566
SEX	−1.19422	0.51555	**−2.31638**	−0.18155
PAR	−0.49581	0.22365	**−2.21690**	−0.07538
PPLI	−0.51624	0.35360	−1.45995	−0.07848

Log-Likelihood: −51.07032 *LRI*: 0.1510

Table 10.20 ML estimates of logit model for $P(E)$ in situation 3

Variable	*Estimate*	*Std. Dev.*	*t-ratio*	*Slope*
CONST	0.49745	1.19447	0.41646	0.05282
GROUP	−1.31165	0.61307	**−2.13948**	−0.13928
JOEX	2.15522	0.59797	**3.60425**	0.22885

Log-Likelihood: −41.34890 *LRI*: 0.1831

Table 10.21 ML estimates of logit model for $P(E)$ in situation 6

Variable	*Estimate*	*Std. Dev.*	*t-ratio*	*Slope*
CONST	2.99789	1.46671	2.04396	0.45221
GROUP	−0.46394	0.50395	−0.92061	−0.06998
SEX	−0.93810	0.51391	**−1.82543**	−0.14151
JOEX	0.36915	0.50127	0.73643	0.05568

Log-Likelihood: −52.34770 *LRI*: 0.0489

The results for situation 2 indicate that gender and parental background are crucial factors influencing the likelihood of a decision in favour of the *equity axiom*. Both SEX and PAR are associated with highly significant partial effects. Again, the probability that a male probant will decide to allocate funds to the worst off (Sub-Saharan Africa) is considerably lower than for a female respondent. Similarly, those who characterize their parental background as being close to a self-employment situation also show less propensity to choose in line with the equity axiom. The covariate PAR is also found to be the best single predictor (highly significant). Moreover, SEX and PAR alone constitute the best two-variable model. It should be emphasized that effects

attributed to SEX and PAR do also hold if one does not control for the position an individual expects to hold in a future income distribution. But adding the PPLI variable has a profound effect on the likelihood of the sample. PPLI was, therefore, included in the final model.

The gender variable which had a considerable partial effect on the probability for the event E in situation 1 and situation 2 does not appear in the final model for situation 3. Here we find a strong negative partial education effect. Controlling for the duration of education, one also finds that respondents with job experience are less inclined to decide in favour of the worst off, the kidney patients.

The results for situation 6 differ from those for situation 1 to 3 with respect to the magnitude of the LR values realized. They are considerably lower. Nevertheless, we identified a final model which contained GROUP, SEX and JOEX. Controlling for the duration of education and job experience, a significant negative effect is attributed to the gender variable at an α level of 0.10. Female respondents decide in favour of basic rights for workers more often than the male respondents do. There is little evidence to support the weak partial effects of JOEX and GROUP.

There is no common set of covariates providing a satisfactory empirical explanation of $P(E)$ across all of the situations. This finding leads us to conjecture that mental processes leading to a decision in line with the equity axiom differ from context to context. According to our results, the only factor that plays a role in the majority of the situations is the gender factor. High partial effects are detected for the SEX variable in situations 1, 2 and 6. In situations 1 and 6 the variable is also the best single explanatory variable, while in situation 2 its partial effect only appears in the presence of PAR and PPLI. The remarkable role attested to the gender variable in the logit models seems to contradict our findings in section 5. There, we failed to produce convincing evidence against the null hypothesis of equal densities for the random variable D for females and males. But one should keep in mind that, in the logit context, we merely model the probability for a particular aspect, namely, the probability for the event that a probant's decision is consistent with the *equity axiom* in the baseline question of a given decision situation. By contrast, the test for the existence of a gender effect in section 5 concerns the equality of the mass distribution over the entire set of elementary events. The fact that the gender effect only emerges in the presence of additional factors underlines the importance of controlling for a variety of probants' characteristics. This is very much in line with the results of Andreoni and Vesterlund (2001) who find that gender has complex effects and interacts with other variables. In addition to the gender variable, GROUP also enters the best model in the majority of situations (situations 1, 2 and 6). Although the GROUP variable has a statistically significant impact coefficient only in situation 3, the impact coefficient is consistently negative in all situations 1, 2 and 3. Finally, notice that parental background, as measured by the PAR-variable, enters the best

model only for situation 2. In view of the fact that this variable can be seen as a complex aggregate variable for which it was difficult to establish prior hypothesis, this result is not very surprising.

Additional evidence on model selection

Here we provide evidence supporting the model selection approach that has been described. For each situation a model has been chosen (cf. Tables 10.18–10.21). Each model is constituted by a subset of the full set of covariates. For each situation, we now estimate the full model and then test whether the coefficients of those variables which are not in the respective best model are *simultaneously* equal to zero using the Wald test and the Likelihood Ratio test (LR).

Under the null hypothesis the Wald test statistic is asymptotically distributed according to a $\chi^2(r)$ distribution. Here, r denotes the number of restrictions imposed. The p-values found in Table 10.22 indicate that there is no evidence against the respective null hypotheses. The deletion of the covariates seems to be justified in each case.

The same type of hypothesis can be tested by an alternative test procedure in each situation. The results for the Likelihood Ratio test procedure are reported below. Note that $LR = -2(ln\, L_{restricted} - ln\, L_{full})$ also is $\chi^2(r)$ distributed if the null hypothesis is true.

The conclusion drawn from Table 10.23 is identical to the one which was produced on the basis of the Wald test.

Table 10.22 Results of the Wald tests

Situation	*Hypothesis* H_0	*df*	*Wald*	*p-value*
1	$\beta_{PAR} = 0 \wedge \beta_{JOEX} = 0$	2	0.4694	0.7908
2	$\beta_{GROUP} = 0 \wedge \beta_{AGE} = 0 \wedge \beta_{JOEX} = 0$	3	1.0972	0.7777
3	$\beta_{SEX} = 0 \wedge \beta_{AGE} = 0 \wedge \beta_{PAR} = 0 \wedge \beta_{PPLI} = 0$	4	1.2346	0.8724
6	$\beta_{AGE} = 0 \wedge \beta_{PAR} = 0 \wedge \beta_{PPLI} = 0$	3	0.3721	0.9460

Table 10.23 Results of the Likelihood Ratio tests

Situation	*Hypothesis* H_0	*df*	LR	*p-value*
1	$\beta_{PAR} = 0 \wedge \beta_{JOEX} = 0$	2	0.4933	0.7814
2	$\beta_{GROUP} = 0 \wedge \beta_{AGE} = 0 \wedge \beta_{JOEX} = 0$	3	1.1219	0.7718
3	$\beta_{SEX} = 0 \wedge \beta_{AGE} = 0 \wedge \beta_{PAR} = 0 \wedge \beta_{PPLI} = 0$	4	1.2542	0.8691
6	$\beta_{AGE} = 0 \wedge \beta_{PAR} = 0 \wedge \beta_{PPLI} = 0$	3	0.3808	0.9442

Probability for reconsidering an initial Rawlsian decision

We now turn to a probant who in the baseline question in a given situation has decided in accordance with the equity axiom, and model the probability of this probant eventually deciding in favour of the better off. This event is given in terms of the elementary events in the following definition.

Definition 4 $S = \{\omega_i \mid w_i \in E \wedge \sum_{j=1}^{4} d_{ij} > 0 \wedge i \in \{0, \ldots, 15\}\}$

An argument identical to the one presented at the beginning of section 6 leads again to a binary response model of the logit type. In trying to explain the probability $P(S)$ we employ the same set of covariates z, the characteristics of the respondents. The models presented in Tables 10.24 to 10.27 were selected by implementing the *all-possible model strategy* outlined above.

Our findings for situation 1 indicate that none of the variables entered into the final model exerts a significant partial effect. It is worth noting though that the 'best' model for $P(S)$ encompasses the variables GROUP, AGE and SEX, that is a subset of the variables constituting the best models for $P(E)$ in situation 1.

For situation 2 the final models for $P(E)$ and $P(S)$ involve identical covariates: SEX, PAR and PPLI. Again, gender is the only factor having a highly significant partial effect on the probability of eventually favouring those who are better off.

In the presence of PPLI and GROUP, both gender and job experience have a significant positive effect on $P(S)$ in situation 3. Males will tend to reconsider their initial decision in favour of the worst off (kidney patients) more often than females. The same holds for those who have progressed to the third year of the advanced business administration programme, relative to those who have just joined the basic programme. Again, the variable SEX entered as the best single explanatory variable with a t-ratio of 2.18. It shows up in best 2, 3 and 4-variate models with slightly increasing standard errors.

Table 10.24 ML estimates of logit model for $P(S)$ in situation 1

Variable	*Estimate*	*Std. dev.*	*t-ratio*	*Slope*
CONST	−0.20900	1.66310	−0.12570	−0.05140
GROUP	0.16450	0.42440	0.38750	0.04050
SEX	0.33570	0.42050	0.79820	0.08260
AGE	−0.03490	0.07280	−0.47890	−0.00860

Log-Likelihood: −65.36690 *LRI*: 0.0064

Table 10.25 ML estimates of logit model for $P(S)$ in situation 2

Variable	*Estimate*	*Std. dev.*	*t-ratio*	*Slope*
CONST	4.11150	1.46630	−2.80400	−0.91010
SEX	1.07560	0.48360	**2.22430**	0.23810
PAR	0.18020	0.17940	1.00430	0.03990
PPLI	0.36420	0.34600	1.05280	0.08060

Log-Likelihood: −50.08840 *LRI*: 0.074

Table 10.26 ML estimates of logit model for $P(S)$ in situation 3

Variable	*Estimate*	*Std. dev.*	*t-ratio*	*Slope*
CONST	2.05790	1.45350	−1.41580	−0.51440
GROUP	0.64040	0.47690	1.34300	0.16010
SEX	0.81580	0.45060	**1.81040**	0.20390
JOEX	−0.86940	0.51740	**−1.68020**	−0.21730
PPLI	0.41250	0.34970	1.17960	0.10310

Log-Likelihood: −57.51270 *LRI*: 0.088

Table 10.27 ML estimates of logit model for $P(S)$ in situation 6

Variable	*Estimate*	*Std. dev.*	*t-ratio*	*Slope*
CONST	−0.91030	1.37350	−0.66270	−0.21970
GROUP	0.31670	0.45840	0.69090	0.07640
PAR	−0.05640	0.15880	−0.35480	−0.01360
JOEX	−0.47760	0.47060	−1.01480	−0.11520
PPLI	0.32230	0.37560	0.85810	0.07780

Log-Likelihood: −58.49240 *LRI*: 0.018

For situation 6, again, the estimates for the final model do not allow for the identification of variables with a significant partial effect. The chosen set of covariates is also different from the one constituting the final model in the case of $P(E)$. In both cases, very low LCR values are observed for the models considered. These observations suggest the need for a respecification of $F(\theta' z)$ and/or the consideration of an alternative set of covariates.

7 Conclusion

Equity judgements are context-dependent. This result follows both from some of the tests carried out in section 5 and from the fact that the set of covariates in the binary response models differs across situations. Moreover, it is in accordance with the conclusion of Gaertner *et al*. (2001).

The individuals participating in our experiments were at two different stages of their education. Those who were at the higher educational level turned out to be significantly more likely to participate until the experiment was completed and to comply with the guidelines provided by the instructor. Such differences in the behaviour of experimental units are often not reported and examined, but need to be considered since they affect the composition of the sample and consequently may limit the generality of the inference. In our case, we found no evidence that non-compliance affected decision patterns.

When testing for differences of the mass distributions of females and males over the entire set of elementary events, we did not find convincing evidence of a gender effect. In the binary response models, however, there were gender effects. Females were invariably closer to the Rawlsian position than were men. Similarly, we found no evidence of educational effects when testing for differences in the mass distributions of individuals at different levels of their education. The results from the estimation of binary response models suggest, however, that individuals at different educational levels may in fact differ when it comes to equity judgements. Our data do not allow us to say more about the nature of the gender and educational effects. On the other hand, the richer results obtained from the estimation of binary response models provide a strong argument for examining the outcomes of experiments by means of state-of-the-art statistical tools.

The five variables GROUP, AGE, SEX, PAR and JOEX share an important common feature. They can all be interpreted as measures reflecting a probant's 'history' up to the point in time when the experiment was carried out. In our binary response models, all these variables, except AGE, appear with a statistically significant impact coefficient in at least one model. This strongly suggests that a probant's history is an important fact that should be taken into account when designing experiments. To obtain more precise information about the impact of the probants' social background, one should in future research try out alternative definitions and measures of the parental background variable.

It is also interesting that we found no significant impact on equity judgements of our only forward-looking variable – expected position in a future income distribution (PPLI). This may be interpreted as being supportive of the view that the probants in our experiment in fact made their equity judgements in a situation that came close to a decision behind the veil of ignorance. In future research, one should include additional forward-looking variables that can be used in more extensive tests of this conclusion.

The fact that variables capturing the history of probants seem to have a strong bearing on social choices opens an interesting perspective. It implies that the equity judgements of the ruling majority in a society can be expected to evolve over time and that equity judgements in societies with different histories can be expected to differ. As indicated earlier, researchers have carried out the experiment reported in the present paper under more or less identical conditions in several countries (Germany, Israel, Latvia, Lithuania and Slovenia). The application of our approaches to testing and econometric modelling, and a careful comparison of the results for each country (culture) has the potential to shed more light on the factors determining distributive judgements.

In our view, the results obtained in this chapter clearly suggest that econometric modelling of probants' decisions in experiments seems to be a fruitful path of research even with the relatively simple models we have used. It might, however, be possible to extract even more information from the data through the use of alternative models. For instance, by considering the six different situations in our experiments as 'cross-sections' in a panel-data framework, it would be possible to take into account possible dependencies between probants' decisions in our six situations. It would also be possible to construct ordered dependent variables measuring the 'degree of adherence to the equity axiom' and estimate ordered logit models. Since this would provide us with a more informative dependent variable, it might give even stronger results, but such modelling strategies would also require stronger assumptions. Alternatively, non-parametric econometric approaches building on weaker assumptions than the binary choice models, should be considered, to check for the robustness of our results.

Appendix: the questionnaire*

Please note that this questionnaire is a part of a larger research project which involves several nations. Please note also that the questionnaire is anonymous. Thank you for your cooperation.

No.:
Group/class:

Version 2

Situation 1:

(o) A small society has received a certain amount of money which can be used either to provide some help and assistance for a handicapped person or to further the education of an intelligent child. The child could receive a good education in languages and in natural sciences, let's say. Let the retarded person be person 1; if the sum of money were used for her support (alternative *x*), she would be able to learn some very basic things, so that at least in certain areas of daily life she

* The questionnaire is reproduced in its original form.

would no longer be totally dependent on the assistance from other people. Let the intelligent child be person 2; the investment into its education represents alternative y. It is not possible to split up the given amount.

Which alternative should be realized in your view, x or y?

(a) Imagine that the sum of money which could be used to help the handicapped person, is so large that, on the other hand, this amount would suffice for the education of not only person 2 but also a second child (person 3) who is even somewhat more intelligent than person 2. Person 3 would, therefore, benefit even a bit more from the education than person 2. Let y be the investment into the education of the two children and let x again stand for the support of the handicapped person.

Would you choose x or y under these conditions?

(b) Imagine that if the money were used to finance alternative y it would be possible to educate still another child (person 4). The reason may simply be "economies of scale" or the fact that a talented teacher will be able to provide a good education for several children simultaneously. Let us assume that all the other characteristics of the situation remain as before.

Which alternative should be picked in your view, x or y?

(c) Add another child to the situation (person 5), who could also receive an instruction in languages and the natural sciences out of the given budget. Everything else remains the same.

Would you want x or y to be realized?

(d_1) If up to this point you have always decided in favour of alternative x, could you imagine a situation, in which you would choose y after all (from the 5th, 6th, 7th, ... intelligent child onward? Or even later?), or would you always decide in favour of the handicapped person, i.e. alternative x?

(d_2) On which criteria did you base your decision? Please give a brief explanation.

Situation 2:

(o) Imagine that due to an unexpectedly large profit of the Federal Reserve (or an unexpectedly large budgetary surplus, if you prefer), Government has the possibility to spend several billion marks (DM) either on environmental protection within its own territory (alternative y) or to spend that amount of money to finance an aid program against hunger in various countries of Subsaharan Africa (alternative x). Given the available amount of money, the environmental program would aim at improving the current situation of the North Sea. This would primarily benefit the fishing industry and, perhaps to a somewhat lesser degree, the people who spend their vacation along the North Sea. Henceforth, these two groups are called "person 2". Those who suffer from famine in Subsaharan Africa are "person 1". Undoubtedly, both the fishermen and the vacationers in this country are, in terms of welfare, better off than the starving people in Africa, independent of whether alternative x or alternative y will be realized. We want to assume that either only x or only y can be realized, not both.

Which alternative should be chosen according to your view, x or y?

(a) Imagine now that the profit of the Federal Reserve (or the budgetary surplus) has turned out to be higher than anticipated originally. On the one hand, the fight against hunger could now be intensified, on the other the environmental program could be extended. The proposal is to improve the quality of the air in the neighbourhood of coal power plants. The group benefiting from this measure will be called "person 3", We shall assume that this group will always be better off than groups 2 and 1 with respect to alternative y, and be definitely better off than group 1 with respect to alternative x. Alternative y again stands for environmental protection and x stands for relief of hunger (both programs would, of course, now be larger due to the higher level of financial resources).

Which alternative should be realized according to your view, x or y?

(b) Assume that it has become clear that "economies of scale" would occur in the environmental program, once alternative y should be realized. We postulate that a program for cleaner water in rivers should also be feasible which would benefit primarily those citizens of the country (group 4) who live close to the rivers (it seems obvious that cleaner water in rivers would, among other things, increase the stock of fish). In other words, not only would groups 2 and 3 benefit from the environmental program but also an additional group. Alternative y again stands for the environmental program and x stands for the aid program for Subsaharan Africa.

Which program should be chosen now, x or y?

(c) Imagine that, given the financial resources, a further enlargement of the environmental program appears realistic. It has, for example, been found out that an additional program aiming at a reduction of traffic noise along the highways would be financially feasible. Through this investment, still another group of people (group 5) would experience an increase in its living conditions. We assume that group 5 is better off than all the other groups under alternative y and that it is at least better off than group 1 under alternative x.

Which alternative should now be realized according to your view, x or y?

(d_1) If up to this point you have always made a decision in favour of alternative x, could you imagine a situation, in which you would choose y after all? And how should y look like in your view, or would you always take a decision in favour of the aid program against hunger?

(d_2) On which criteria did you base your decision? Please give a brief explanation.

Situation 3:

(o) Imagine a country that has a severe shortage of western currencies. The governing body of this country has the possibility to purchase on the world market either a certain number of badly needed dialysis machines (alternative x) that cannot be produced within the country, or a certain quantity of vitamin pills as well as tropical fruit (alternative y). This quantity would only be enough to satisfy the urgent needs of a relatively small group of persons. The realization of both alternatives together or a combination of both alternatives to some extent is assumed to be infeasible. The group of people suffering from kidney problems is group 1, the group of people benefiting from the import of vitamins and fruits is group 2. There is unanimous agreement in the country that all pregnant women should

make up group 2. It is also unanimously agreed that the persons with kidney trouble are clearly worse off than the expectant mothers.

Which alternative should be realized in your view, x or y?

(a) Imagine now that the world market price for vitamin pills and tropical fruit has fallen. If alternative y were realized it would be possible to provide not only the expectant mothers, but also all the country's babies and toddlers (group 3) with the needed vitamins. The price of dialysis machines is assumed to rest unchanged, however. The welfare levels of groups 2 and 3 are clearly higher than the level of group 1 both under y and under x.

Would you choose alternative x or alternative y?

(b) Let us imagine that there is a further decline in the world market price for vitamin pills and tropical fruit so that it turns out that under the given amount of western currencies the country's adolescents (group 4) could also be provided with vitamins if alternative y were chosen.

Which alternative should be chosen, x or y?

(c) The world market price of vitamin pills and tropical fruit declines once more so that under alternative y the given amount of western currencies would now suffice to provide those workers of the country who are engaged in physical labour (group 5) with the needed vitamins. Clearly, these workers are better off, no matter whether they receive the vitamins or not, than the group of persons who suffer from kidney problems.

Which of the two alternatives, x or y, should now be chosen?

(d_1) If up to this point you have always made a decision in favour of alternative x, could you imagine a situation, in which you would choose y after all? And how should y look like in your view, or would you always take a decision in favour of x?

(d_2) On which criteria did you base your decision? Please give a brief explanation.

Situation 4:

(o) Imagine a country which has a severe shortage of western currencies. The governing body of this country has the possibility to purchase on the world market either a certain number of badly needed dialysis machines (alternative x) that cannot be produced within the country, or a certain quantity of expensive wines from the Bordeaux region (alternative y) that are desired by certain segments of the society. The realization of both alternatives together or a combination of both alternatives to some extent is assumed to be infeasible. It is hypothesized that the wines from Bordeaux have such a high price that they could only be purchased by a small group of relatively well-off citizens (group 2). The group of people suffering from kidney problems is group 1. It goes without saying that due to their illness, the dialysis patients are worse off than the potential buyers of expensive French wines. Alternative x refers to the import of dialysis machines and y refers to the import of wines from Bordeaux.

Which of the two alternatives should be chosen according to your view, x or y?

(a) Imagine that the price of Bordeaux wines has fallen so that a second group within society (group 3) would be able to purchase these wines if alternative y were

realized. Clearly, the quantity of imported wines could be increased due to the lower market price. The price of dialysis machines is supposed to remain constant.

Which alternative should be selected now, x or y?

(b) Let us assume that a further decrease in price of the Bordeaux wines has occurred so that under the given amount of western currencies an even larger quantity of wines could be imported now. Therefore, due to the lower price per bottle, yet another group within society (group 4) could become a buyer of those wines.

Which alternative should now be realized, x or y?

(c) The price of wines from Bordeaux is supposed to fall once more so that, again, another group within society (group 5) would be put in a position to purchase these wines if alternative y were realized.

Would you choose x or y in this situation?

(d_1) If up to this point you have always made a decision in favour of alternative x, could you imagine a situation, in which you would choose y after all? And how should y look like in your view, or would you always take a decision in favour of x?

(d_2) On which criteria did you base your decision? Please give a brief explanation.

Situation 5:

(o) Once again, imagine a country with a severe shortage of western currencies. The governing body of this country has the possibility either to purchase on the world market a certain amount of inexpensive clothing (alternative x) which would allow the more needy segments of society (group 1) to significantly improve their welfare level, or to import a certain quantity of expensive wines from the Bordeaux region (alternative y) that a small group of rather well-to-do citizens of the country (group 2) would like to acquire. The realization of both alternatives together or a combination of both alternatives to some extent is assumed to be infeasible.

Which alternative should be chosen according to your view, x or y?

(a) Imagine that the price of Bordeaux wines has fallen so that a second group within society (group 3) would be able to purchase these wines if alternative y were realized. This additional group 3 is supposed to be better off in terms of welfare than group 1. We assume that the price of inexpensive clothing remains the same so that the quantity of imports would not change, should x be realized.

Should x or y be chosen?

(b) Let us assume that a further decrease in price of the Bordeaux wines has occurred so that with the given amount of western currencies an even larger quantity of wines could be imported now. Therefore, due to the lower price per bottle, yet another group within society (group 4) could become a buyer of these wines.

Which of the alternatives x or y should now be realized?

(c) The price of wines from Bordeaux is supposed to fall once more so that, again, another group within society (group 5) would be put in a position to purchase these wines if alternative y were realized.

Would you choose x or y in this situation?

(d_1) If up to this point you have always made a decision in favour of alternative x, could you imagine a situation, in which you would choose y after all? And how should y look like in your view, or would you always take a decision in favour of x?

(d_2) On which criteria did you base your decision? Please give a brief explanation.

Situation 6:

(o) Imagine a country which had been totally run down economically by a long-lasting dictatorship. Finally, the country could get rid of this dictatorship. Furthermore, imagine that an international bank group is offering a rather large loan (under very favourable conditions of repayment) to this country for economic reconstruction (alternative y). However, the consortium declares that the prerequisite for this loan should be that the employees in the country be granted neither a right to strike nor the free choice of occupation. This precondition would remain valid for the foreseeable future. If the new Government were unwilling to enforce this curtailment of individual rights, no loan would be offered, and, therefore, the country would have to pull itself up by its bootstraps (alternative x). In that case, the country would, of course, have the option to reinstall the right to strike and other basic rights, a measure which had been promised to the citizens of the country after the fall of the dictatorship. If the bank loan were granted, the large enterprises (group 2) would be the first to experience an economic recovery. The workers and employees in the firms (group 1) would be hard hit by the restriction of basic rights. Also, their economic situation would be worse than that of the people in charge of the large enterprises.

What should the country do in your view, should it decide in favour of y or x?

(a) Imagine that the initial situation were to undergo the following modification: The loan which is offered would have such a large volume that an additional group of the population, the self-employed persons with a small or middle-sized business activity, let's say, would benefit from the financial aid (group 3). Let this alternative again be denoted by y. Alternative x remains as before.

Should the country choose x or y?

(b) Imagine again a change of the initial situation: The bank loan offered is so large that under alternative y still another group of the population, the civil servants, let's say, would realize larger economic benefits (group 4). Alternative x remains unchanged.

Which alternative should now be picked by the country?

(c) A further variation: we shall assume that still another group within the population, the retired members of society (group 5), would experience an improvement of their economic situation under alternative y. Alternative x remains unchanged.

Which alternative should now be chosen according to your view, x or y?

(d_1) If up to this point you have always made a decision in favour of alternative x, could you imagine a situation, in which you would choose y after all? And how should y look like in your view, or would you always take a decision in favour of x?

(d_2) If right from the beginning or "fairly quickly" you took a decision in favour of y, what kind of reasons or motives could have moved you to decide in favour of x? A further confinement of basic rights such as a law against founding political parties, the introduction of censorship of the press, TV etc. or a confinement of religious freedom? Or something else? Please give a brief explanation.

Demographic characteristics

(1) sex: (1 = female, 2 = male)
(2) age:
(3) Which of the following categories describes best the professional status of the family in which you grew up?

(1 = unskilled worker, 2 = skilled worker, 3 = craftsman,
4 = employee or civil servant in the public sector,
5 = employee in the private sector,
6 = self-employed)

(4) subject of study: (1 = business administration, 2 = mathematics, 3 = economics, 4 = other)
(5) Were you employed before starting with your studies?
(1 = yes, 2 = no)
(6) How many percent of the citizens of your country do you expect to have, in the year 2010, a net-income lower than your own?
(1 = 5%, 2 = 25%, 3 = 50%, 4 = 75%, 5 = 95%)

References

Agarwal, B. (2000) 'Conceptualizing Environmental Collective Action: Why Gender Matters', *Cambridge Journal of Economics*, vol. 24, pp. 283–310.

Amiel, Y. and F. Cowell (2000) 'Attitudes towards Risk and Inequality', A Questionnaire-experimental Approach; Discussion Paper no. DARP 56, London School of Economics.

Andreoni, J. and L. Vesterlund (2001) 'Which is the Fair Sex? Gender Differences in Altruism', *the Quarterly Journal of Economics*, vol. 116, pp. 293–312.

Bolton, G.E. and E. Katok (1995) 'An Experimental Test for Gender Differences in Beneficient Behavior', *Economics Letters*, vol. 48, pp. 287–92.

Brown-Kruse, J. and D. Hummels (1993) 'Gender Effects in Laboratory Public Goods Contribution', *Journal of Economic Behavior and Organisation*, vol. 22, pp. 255–67.

Cadsby, C.B. and E. Maynes (1998) 'Choosing between a Socially Efficient and Free-riding Equilibrium: Nurses versus Economics and Business Students', *Journal of Economic Behavior and Organization*, vol. 37, pp. 183–92.

Carter, J.R. and M.D. Irons (1991) 'Are Economists Different, and If So, Why?', *Journal of Economic Perspectives*, vol. 5, pp. 171–77.

Davidson, R. and J.G. MacKinnon (1993) *Estimation and Inference in Econometrics* (Oxford: Oxford University Press).

Deschamps, R. and R. Gevers (1978) 'Leximin and Utalitarian Rules: A Joint Characterization', *Journal of Economic Theory*, vol. 17, pp. 143–63.

Eckel, C.C. and P.J. Grossman (1998) 'Are Women Less Selfish than Men? Evidence from Dictator Experiments', *Economic Journal*, vol. 108(2), pp. 726–35.

Frank, B., and G.G. Schultze (2000) 'Does Economics Make Citizens Corrupt?', *Journal of Economic Behavior and Organization*, vol. 43, pp. 101–13.
Frank, R.H., T. Gilovich, and D. Regan (1993a) 'Do Economists Make Bad Citizens?', *Journal of Economic Perspectives*, vol. 10, pp. 187–92.
—— (1993b) 'Does Studying Economics Inhibit Cooperation?', *Journal of Economic Perspectives*, vol. 2, pp. 159–71.
Frohlich, N., J.A. Oppenheimer, and Ch. Eavey (1987a) 'Choices of Principles of Distributive Justice in Experimental Groups', *American Journal of Political Science*, vol. 31, pp. 606–36.
—— (1987b) 'Laboratory Results on Rawls's Principle of Distributive Justice', *British Journal of Political Science*, vol. 17, pp. 1–21.
Gaertner, W. (1992) 'Distributive Judgements' in W. Gaertner and M. Klemisch-Ahlert (eds), *Social Choice and Bargaining Perspectives on Distributive Justice* (Springer Verlag, Heidelberg, Berlin, New York), pp. 17–59.
—— (1994) 'Distributive Justice: Theoretical Foundations and Empirical Findings', *European Economic Review*, vol. 38, pp. 711–20.
Gaertner, W., and J. Jungeilges (2002) 'Evaluations via Extended Orderings: Empirical Findings From West and East', *Social Choice and Welfare*, vol. 19, pp. 29–55.
Gaertner, W., J. Jungeilges, and R. Neck (2001) 'Cross-cultural Equity Evaluations: A Questionnaire-experimental Approach', *European Economic Review*, vol. 45, pp. 953–63.
Gilligan, C. (1982) *In a Different Voice: Psychological Theory and Women's Development* (Cambridge, MA: Harvard University Press).
Gray, J. (1991) *Men Are from Mars, Women Are from Venus: A Practical Guide for Improving Communication and Getting What You Want in Your Relationship* (New York: Harper Collins).
Green, W.H. (1990) *Econometric Analysis* (London: Macmillan).
Hargreaves-Heap, S., and M. Hollis (1987) 'Epistemological Issues in Economics' J. Eatwell, M. Milgate, and P. Newman (eds) In *The New Palgrave* (London: Macmillan) vol. 2, pp. 166–8.
Hogg, R.V., and A.T. Craig (1978) *Introduction to Mathematical Statistics* (New York: Macmillan).
Jacobsen, D.I. (2001) 'Higher Education as an Arena for Political Socialisation: Myth or Reality?', *Scandinavian Political Studies*, vol. 24, pp. 351–68.
Jungeilges, J., and T. Theisen (2003) 'An econometric examination of equity judgements elicited through experiments', Working Paper, Department of Economics and Business Administration, Agder University College, Kristiansand, Norway.
Laband, D.N., and R.O. Beil (1999) 'Are Economists More Selfish than Other "Social" Scientists? , *Public Choice*, vol. 100, pp. 85–101.
Manski, C.F. (2002) 'Identification of Decision Rules in Experiments on Simple Games of Proposal and Response', *European Economic Review*, vol. 46, pp. 880–89.
Marwell, G., and R.E. Ames (1981) 'Economists Free Ride, Does Anyone Else? Experiments on the Provision of Public Goods', *Journal of Public Economics*, vol. 15, pp. 295–310.
Mason, C.F., and O.R. Philips (1991) 'The Role of Gender in a Non-cooperative Game', *Journal of Economic Behavior and Organization*, vol. 15(2), pp. 215–35.
Novell, C., and S. Tinkler (1994) 'The Influence of Gender on the Provision of a Public Good', *Journal of Economic Behavior and Organisation*, vol. 25, pp. 25–36.

Ortmann, A., and L.K. Tichy (1999) 'Gender Differences in the Laboratory: Evidence from Prisoner's Dilemma Games', *Journal of Economic Behavior and Organization*, vol. 39(3), pp. 327–39.

Rapoport, A., and A. Chammah (1965) *Prisoner's Dilemma: A Study in Conflict and Cooperation* (Ann Arbor, MI: University of Michigan Press).

Rawls, J. (1971) *A Theory of Justice* (Cambridge, MA: Harvard University Press).

Rosener, J.B. (1990) 'Why Women Lead', *Harvard Business Review*, vol. 68(5), pp. 119–25.

Sell, J., W.L. Griffith, and R.K. Wilson (1993) 'Are Women More Cooperative than Men in Social Dilemmas?', *Social Psychology Quarterly*, vol. 56, pp. 211–22.

Sell, J. (1997) 'Gender, Strategies, and Contributions to Public Goods', *Social Psychology Quarterly*, vol. 60, pp. 252–65.

Selten, R., and A. Ockenfels (1998) 'An Experimental Solidarity Game', *Journal of Economic Behaviour and Organization*, vol. 34, pp. 517–39.

Solnick, S.J. (2001) 'Gender differences in the ultimatum game', *Economic Enquiry*, vol. 39, pp. 711–20.

Tannen, D. (2001) *You Just Don't Understand Women and Men in Conversation* (New York: Quill).

Yaari, M.E., and M. Bar-Hillel (1984) 'On Dividing Justly', *Social Choice and Welfare*, vol. 1, pp. 1–24.

Yetzer, A.M., R.S. Goldfarb, and P.J. Poppen (1996) 'Does Studying Economics Discourage Cooperation? Watch What We Do, Not What We Say or How We Play', *Journal of Economic Perspectives*, vol. 105, pp. 177–86.

11

Groups, Commons and Regulations: Experiments with Villagers and Students in Colombia*

Juan Camilo Cardenas
Facultad de Economia, Centro de Estudios sobre Desarrollo Económico (CEDE), Universidad de Los Andes, Colombia

1 Introduction

Group externalities imply a situation where individual and group interests are not aligned and therefore require the design of rules or institutions that correct the failure in order to improve social outcomes. Public goods, team work, the use of natural resources under joint access, or any pollution problems, are examples of such potential divergence between individual and group incentives. Institutional corrections can come exogenously from a regulatory state that brings in command and control or incentive mechanisms (pecuniary or not-pecuniary), or that reassigns property rights to correct the failure. But solutions can also emerge endogenously from the group, through self-governed institutions, with similar mechanisms of material or non-material incentives, as well as social norms or conventions.

To solve the dilemma, that is, to induce individuals to make decisions that are both individually rational and socially efficient, we need to set up institutional mechanisms that are cost-effective. But the design and

* This research was made possible through a Research and Writing Grant from the John D. and Catherine T. Macarthur Foundation and a grant from the Network on Social Norms and Preferences headed by Herbert Gintis and Robert Boyd. I thank them for their support. Thanks are also due to Ernst Fehr in particular for providing important insights for the experimental design; to the Santa Fe Institute for an International Fellowship, and to the School of Environmental and Rural Studies at Javeriana University (Colombia) during my field work. I also thank Maria Claudia Lopez, Pablo Ramos and Ana María Roldán for assistance in the field and in processing data. Diana Maya provided important inputs in the analysis of qualitative data. I am also grateful to Maria Alejandra Velez and the editors of this volume (in particular Bina Agarwal) for their valuable comments that helped improve this manuscript.

enforcement of such mechanisms are likely to involve a range of costs, such as those associated with the design of the instrument, with the gathering of information about compliance and non-compliance, and with punishing violators and/or rewarding those complying with material or non-material incentives. But if these costs are sufficiently high, we cannot guarantee full compliance when the regulator – whether from the group or external – has incomplete information, and there is thus room for opportunistic behaviour on the part of the agents.

Nevertheless, as we observe from the experimental results reported here, and from the expanding literature on other-regarding preferences (Bowles, 1998; Gintis, 2000; Camerer and Fehr, 2004; Fehr and Schmidt, 1999), there are still possibilities of pro-social or group-oriented behaviour, even in the absence of any enforceable mechanism, which can help solve the conflict between individual and group interest. Basically, without perfectly monitored and cost-effective enforcement, the ability of any partially enforceable institution to solve the dilemma will ultimately depend on individual willingness to follow the social norm. This would solve the problem of coordination failure, given the asymmetries of information between the regulator and the group members. This is true for both an external regulator (for example the state) and an endogenous one (for example a community organization). Both can only partially observe the level of cooperation in collective action, and the level of compliance of any new mechanism that is introduced.

The behavioural focus here is twofold. On the one hand, this chapter will explore what choices individuals (as regulated agents) make when an external regulation is introduced to solve the problem of coordination failure. On the other hand, it will examine how individuals respond when asked, through a voting mechanism, about their preferences towards the application of such mechanisms by an external regulator, for example the state.

This chapter seeks to make a small contribution by testing some of the behavioural hypotheses under different regulations, with an experimental design tested both in the field and in the university laboratory. Although the experimental literature has explored several behavioural phenomena associated with monitoring and sanctioning, many of these have focused on sanctioning by the same group members, by allowing players to reduce or increase the income of other players, depending on their cooperative behaviour. The mechanisms through which individuals are willing to punish free-riders, at a personal cost, has been studied extensively, including by Carpenter (2004) and Carpenter and Mathews (2004) who suggest that there is a demand for punishment in people's utility function. Andreoni *et al*. (2003) also study the individual and combined effects of rewards and sanctions, showing that in combination they can be effective in inducing individuals to increase social efficiency in a proposer–responder game. But again, it is the players themselves who assume the cost of rewards and sanctions, rather than an external agent.

The path taken here, however, departs from these approaches.[1] First, we wanted to study the behavioural consequences of external regulators introducing rules to induce agents to become more cooperative. Second, we wanted to introduce the problem of imperfect monitoring in which the regulator cannot observe or sanction more than a fraction of the players. These conditions seem to resemble many instances in which government agencies can only partially monitor the behaviour of agents in order to sanction those not complying with a socially-oriented norm.

Besides finding behavioural differences between college students and people in the field, which should be of interest to those using experimental methods in general, the results suggest that even if a majority of players vote against the rules to enforce a socially-optimum outcome, they are still willing to cooperate and reduce overextraction, as suggested by the regulations tested. Further, the data do not entirely support the symmetric Nash equilibrium predicted by a model of the expected costs of different regulations. Players do not respond substantially to changes in the size of the penalty. Also, differences across the sites we studied might explain attitudes towards the rules, the rulers and the compliance levels.

Section 2 describes the village context where the field experiments were conducted; section 3 presents the theoretical model and experimental designs; section 4 presents the results; and the concluding section 5 outlines some implications.

2 The field context

The experiments and fieldwork were conducted during 2000–03 in 10 different sites across Colombia. For the entire sample of sessions we randomly distributed different regulations and payoff structures among almost 20 different experimental treatments. The treatments reported here were run in five of these villages after making sure that within each village at least three sessions of exactly the same treatment were conducted. In all sites we recruited participants that had some dependence on a natural resource for which there was joint access, although individually such dependence could vary within a site. The five sites, in alphabetical order, for this sample are described below.[2]

In the dam reservoir of Neusa, about 100 km north of Bogotá in the Andean region, a group of farmers have been engaged in trout fishing mainly to sell in the local markets and restaurants, keeping a small proportion for self-consumption. They have been experiencing a fall in the resource stock due to overfishing and competition from sport fishermen. The regional environmental authority has been introducing stronger regulations on catch size and quotas, as well as reintroducing trout populations, given that trout cannot reproduce naturally in these settings. About 40 farming households directly participate in this activity while farming their own land

alongside, and dozens come during the week and weekends for sport fishing.

In the Island of Providencia, in the Caribbean coast of Colombia, we invited those who participate in two main economic activities of extraction. One group we invited was of fishermen that catch several species in the coral reefs and coastal waters for selling to passing boats and local markets, as well as for self-consumption. The other group we invited regularly gathers crabs that circulate between the inland and the beach. They sell these to the local restaurants that cater to tourists and also gather some for self-consumption. In both cases, fishing and crab catching, there are regulations that are partially enforced by the regional authority. In addition, there are a few informal and community agreed rules for self-monitoring catch sizes and location rotations, by seasons.

In the southern Pacific coast of the country we find the National Natural Park of Sanquianga, composed mostly of mangrove forest that provides a multiplicity of goods and services to people spread in small settlements constituted of a few dozen or a few hundred households. They allot most of their time to fishing and gathering other products from the coastal areas such as fish, clams, firewood, crabs, and so on. We particularly recruited people that extract a mollusc or clam (*anadara tuberculosa*) which provides food and income but which has also been suffering from overextraction and a declining market over the years. The officers of the National Park regulate the area through a few weakly-enforced rules issued by the national and regional environmental authorities and a few agreements with communities and organized groups. They seek to limit catch sizes, net types and sizes, and to enforce the closed season for some of the resources by biological seasons.

In the community of Tabio, about 50 km northwest of Bogotá and also in the Andean mountains, a few hundred farmers and rural households depend on the natural supply of water from their local watershed. A remaining fraction of households extract firewood from these forests for cooking. The supply of water depends highly on the natural vegetation, and the local municipality and regional authorities try to enforce a set of rules for the buffer zones along the watershed, which is mostly located within private property.

In the Andean forest about 150 km west of Bogotá, at an elevation of around 1,500 meters we find the village of La Vega where farmers share a watershed and need to make decisions about land use, afforestation and firewood extraction along with decisions relating to their agricultural and livestock practices, and especially regarding coffee production which requires water for irrigation and coffee beans processing. These activities create organic emissions downstream. The supply of water is highly dependent on the management of the natural forest, especially upstream where the water springs are located. The local committee of coffee growers has been working with local authorities and farmers' associations to undertake

training and technological transfer activities to make a better use of the forested area around the watershed and improve the quality of resources, increase and sustain the supply of firewood and continue afforestation by planting native species.

In all five cases we invited adults (18 years and above) to participate in the experiments. No two people from the same household were allowed to participate in the same session. We also tried to recruit people from all socio-economic levels in the villages. About half of the total participants were females, but depending on the resource and location, our sample could have a selection bias: for instance, in the case of Providencia most fishery activities are conducted by men and most crab gathering by women. However, no significant gender effect was found in experimental behaviour.

We found some common patterns across the sites. We have rural households that depend in some degree on a joint access natural resource that is used for self-consumption and/or market sale. We observe that households must allocate their resources (labour, land, equipment) into private activities and extractive activities. This leads to an inherent paradox: on the one hand extracting more of the resource enhances individual well-being. On the other hand at an aggregate level, more extraction produces negative externalities for the community as a whole. In all cases there is a regulatory agency that partially enforces some rules that limit extraction or use of the resource, but we observe that the monitoring and enforcement costs affect compliance.

3 The experimental and field approach

There is an extensive theoretical, empirical and experimental literature on the study of self-governed solutions to dilemmas such as the 'tragedy of the commons'. Much of this literature has been catalyzed by Hardin's postulates in the late 1960s. Writings by Berkes (1989), Ostrom (1990), Ostrom, Gardner and Walker (1994), and Baland and Platteau (1996) are seminal works on the topic, that have helped us qualify the conditions that seem to predict or reject the tragedy. More recent work which combines and contrasts experimental, theoretical and field-based analysis provides a wide set of conditions under which individuals may confirm or reject the original tragedy scenario (Ostrom *et al.*, 2002). Many of these factors are related with technology in the group production function, or with the relative marginal returns from group production, or with the private or exit options of group members, or with the surrounding conditions that provide an ecological or cultural environment for the emergence of certain norms of cooperative behaviour.

The experimental literature on voluntary contributions to a public good, or the extraction of a common-pool resource suggests that individuals respond with a wider set of strategies than those predicted from a simple game-theoretical solution based on the maximization of net personal material gains (Ledyard, 1995; Ostrom, Gardner and Walker, 1994). In fact some are

willing to cooperate even in one-shot anonymous settings. However, in the absence of any institution that coordinates individual actions, the tendency in a repeated game is for free-riding to increase over time and lead to less socially-optimal equilibria. The most plausible explanation for this is that when players see others become more opportunistic it induces a reciprocal response, which decreases social efficiency over time.

Regulatory agencies can design various mechanisms to induce a change in the incentives of the players and so improve group outcome or decrease over-extraction of a common-pool resource. Assuming that the regulator wants to maximize the group outcome,[3] it can introduce a regulation that imposes a cost on individuals who act in ways that reduces the group outcome. The severity of the cost can be made variable by changing the level of monitoring or enforcement, the level of the fine or penalty, or the way the regulated players participate in the decision to impose or not to impose such a mechanism.

As a baseline scenario, the regulator can assume that under a non-regulated and non-cooperative game, players make their decisions based on a Nash strategy that maximizes their individual material payoffs. The regulator can, for instance, impose a cost on certain decisions so that the new material payoffs (that include the expected cost of the regulation) induce a new Nash strategy that achieves more efficient outcomes for the group. However, unless the regulator can fully enforce the new rule by monitoring all players, there might be risks for opportunistic behaviour by players that overextract the commons incorporating into their calculations the probability of not being inspected and therefore benefit from the compliance of others. As the fraction of players following this opportunistic strategy grows, the aggregate extraction of the resource brings the common-pool to socially inefficient levels.

One problem with regulations designed at shifting behaviour towards optimal group outcomes is that they assume players behave as predicted by the *homo economicus* model of behaviour, in which individuals follow an optimizing strategy based solely on their own material payoffs, including the expected costs of the regulation. However, economic decisions are also affected by other factors, such as emotions, social norms, reciprocity, inequality aversion, lack of complete information and lack of perfect calculating capabilities, among others, as behavioural and experimental economists have demonstrated (Kahneman and Tversky, 2000; Rabin, 1998, 2002; Fehr and Schmidt, 1999; Loewenstein, 1999).

In Cardenas *et al*. (2000) we tested such hypotheses by conducting a series of experiments in the field with several groups of eight villagers who in their daily lives face the problem of joint access to a natural resource. First, the results showed that players did not choose their best Nash response during the first stage of a repeated number of rounds, where no regulation or coordination was allowed among the players. Second, and more interestingly,

during the second stage when a new rule was introduced that imposed a fine, proportional to the level of extraction, with a probability of 1/16 to be enforced for each of the eight players, the effectiveness of the regulation was very high in the first rounds. But with subsequent rounds group behaviour became increasingly individualistic, lowering the levels of cooperation even below those achieved during the first stage with no regulation at all.

These factors that affect behaviour should also be taken into account when designing regulations aimed at guiding or shifting rational individual behaviour away from socially-inefficient decisions. The experimental literature has explored the effects of regulations and explicit economic incentives in common-pool public goods and other social dilemmas settings, but most use university students as their main subjects. Further, much of the experimental work on the role of monitoring and sanctioning has focused on mechanisms that are endogenous to the group, that is, the group members are assumed to take on the cost of monitoring or even reducing the penalization of free-riders.[4]

However, there is surprisingly little work on situations where external regulators enforce rules aimed at correcting group externality, despite the fact that in many field settings where groups depend on a common-pool resource it is usual to have an external authority that devises and implements regulations based on incentives, or on command and control. Furthermore, although experiments on regulations and incentives have now been conducted for some time, barring those undertaken in 1998 by Cardenas *et al*. (2000), few such experiments have been done in field settings where the subjects in real life face the theoretical problem in question.[5]

This chapter explores individual decision-making under such economic settings by bringing the experimental laboratory to the field and compares the results with the experiments involving students. In particular, I focus on changes in individual decisions concerning resource extraction and their effects, after introducing external regulatory mechanisms aimed at solving the dilemma. I also examine individual preferences for such external regulatory mechanisms.

The design is based on a simple model of a group of natural-resource users who face a group externality, to test the effects of regulations and rule enforcement in these dilemmas. The experimental situation is framed as one where a group of five people can extract resources from the same source to which they have joint access (see Appendix for protocols and instructions read to the participants). Individuals derive part of their payoffs (direct benefits) from putting effort into extracting a resource (such as firewood). They also see their payoffs (indirect benefits) reduced in terms of, say, a fall in water quality or a loss of biodiversity from the negative externality that occurs as the group's aggregate extraction increases, because of overextraction from the common pool. In each of our experimental sessions, over 20 rounds, five participants must decide the level of resource extraction, where individual extraction increases payoffs but in aggregate terms the extraction reduces

your own and the others' payoffs. The design based on repeated rounds will allow players to understand this incentive structure and form conjectures of what the rest of the players in their group would do within the particular set of rules.

For the last 10 rounds we introduced a rule aimed at achieving the social optimum by using an explicit material cost to individual overextraction, but the monitoring of the rule was stochastic and incomplete, that is, only one of the players was monitored and actually faced the fine or penalty if extraction was above the permitted level. We altered the penalty, as well as the way the rule was decided (mandatory or through group voting and applied it only if a majority approved). We then compared these treatments with a baseline one where for the 20 rounds the participants were not allowed to communicate, nor did they face any kind of regulation.

All sessions in the field and the university were hand-run, using pencil and paper. Each session took about three hours. This included the time taken for reading the protocols (see Appendix), running two or three practice rounds, undertaking two stages of 20 rounds, and calculating the earnings by subject while they filled out a short exit survey and other demographic details.

4 Field application of an experimental design

Between 2000 and 2003 we conducted a long series of economic experiments in the five rural sites mentioned in section 2, with 320 villagers that depended on common-pool resources on a daily basis. We also conducted the same experiments with 105 college students in order to compare and derive some lessons regarding the use of these methods both in the lab and in the field.

The non-cooperative game

Assume n players that benefit from the forest, and who each have the same maximum labour endowment of e units to allocate in extracting a particular resource. Player i's level of extraction x_i, with $0 \leq x_i \leq e$, increases her payoffs at a decreasing rate (direct extraction benefits $= ax_i - \frac{1}{2}bx_i^2, a, b > 0$), while the aggregate extraction by the n players $\sum x_i (i = 1, n)$, reduces i's payoffs (indirect benefits $= \sum(e - x_i), > 0$). The externality can also be described as a public good benefit from conservation, that is lack of extraction. Thus, player i's payoffs are:

$$\pi_i = \left(ax_i - \frac{1}{2}bx_i^2\right) + \alpha \sum (e - x_i) \tag{1}$$

For n players, and assuming symmetric endowments for all, e, we can rewrite (1) as:

$$\pi_i = ax_i - \frac{1}{2}bx_i^2 + \alpha ne - \alpha \sum x_i \tag{2}$$

If player i chooses x_i to maximize Π_i, the first-order conditions that produce the optimal level of extraction x_i^{nash} are $\partial\pi_i/\partial x_i = a - bx_i - \alpha = 0$, which requires that:

$$x_i^{\text{nash}} = (a - \alpha)/b, \quad \text{for } 0 \leq x_i \leq e \tag{3}$$

For example, suppose values of $e = 8$, $a = 60$, $b = 5$, and $\alpha = 20$, $x_i^{\text{nash}} = (a - \alpha)/b = 8$. From the standpoint of any player it would be a Nash best response – and a dominant strategy – to allocate maximum labour for extracting from the forest. Below we choose parameters that allow us to create the environment of incentives we need for our experimental design.

To produce the socially efficient outcome, we maximize the aggregate payoffs (4) and calculate the optimal level of extraction for the individuals, x_i^{so}.

$$W = \sum \pi_i = \sum a x_i - \frac{1}{2} b \sum x_i^2 + \alpha n^2 e - \alpha n \sum x_i \tag{4}$$

The first-order conditions, $\partial W/\partial x_i = a - bx_i - \alpha n = 0$, which requires that:

$$x_i^{\text{so}} = (a - \alpha n)/b, \quad \text{for } 0 \leq x_i \leq e \tag{5}$$

For the same parameters as above, and for a group of $n = 5$ players, such a solution would require each player to allocate $(60 - 5 * 20)/5 < 0$. Since x_i takes only non-negative values, for framing the experimental design we have a corner solution at $x_i^{\text{so}} = 0$, that is players should not allocate labour into extraction to produce the socially efficient outcome. We have eliminated in the table the option of $x_i^{\text{so}} = 0$ to avoid possible conflicts in conducting these experiments in the field. Previous experiments and pre-testing exercises suggest that there is a strong aversion towards prohibition of resource use that could create problems with our participants when conducting the experiments this way. Interior solutions with non-dominant strategies, such as used in Ostrom, Gardner and Walker (1994) and Cardenas *et al.* (2000), are another alternative, but here too we have decided to maintain corner solutions in order to have a design with a dominant strategy.

The payoffs for these parameter values are given in the Appendix as the payoffs table that was used in the field for the experiments. The values in the cells are the earnings in one round. This table was handed to the five participants in each group. In each round, each person should decide the level of extraction between one and eight units, and the earnings can then be calculated if we know the total group effort expended in extracting, by looking at the respective column and row in the table. For instance, if player i follows the group maximizing strategy $x = 1$, individual earnings are \$745 given that i's extraction is 1 unit and 'their extraction' is 4 units. If all players

follow the Nash dominant strategy, $x = 8$, each player will obtain \$320 in that round.

External regulation of individual behaviour

Following the analysis of external regulation in Cardenas *et al*. (2000) we introduce a regulation in the form of a penalty f (cost), externally imposed with certain probability p, to individual allocations above that required for the social optimal solution. By varying the probability of inspections times the fine or penalty we will test weak and strong enforcement levels. The treatments will also include applying the regulation exogenously (by the monitor), and endogenously (voted by the group).

Following from equation (2), the new expected payoff function for a player i is then:

$$\pi_i = ax_i - \frac{1}{2}bx_i^2 + \alpha ne - \alpha \sum x_i - pf(x_i - x_i^{so}) \tag{6}$$

which yields the first-order conditions: $\partial \pi_i / \partial x_i = a - bx_i - \alpha - pf = 0$, which requires that:

$$x_i^{\text{nash-REG}} = (a - \alpha - pf)/b, \quad \text{for } 0 \leq x_i \leq e \tag{7}$$

Suppose that for the weak enforcement case the probability of inspection is $p = 1/5$ and the fine $f = 50$. Here the best response, based in the expected cost of the regulation, should be $x_i^{\text{nash-REG}} = 6$ units, which would only partially increase earnings. In the case of a strong regulation for a fine of $f = 175$ points, we get $x_i^{\text{nash-REG}} = 1$ unit, which would achieve the socially optimal solution by aligning the best response function with the socially optimal condition.

Field experiment design

The experiments were all conducted for groups of five players, in a finite repeated game of 20 rounds, divided into two stages of 10 rounds each. The players' payoffs, according to the above model and the payoffs table included in the Appendix, increase with the individual's level of appropriation but decrease with the aggregate level of use by the group, following the incentives structure of any group externality dilemma. Throughout the game all decisions are made individually and privately and only the groups' outcome is publicly announced in each round. The first stage (rounds 1–10), for all groups, will be under a baseline treatment as a non-cooperative game, where each subject decides individually and privately her level of appropriation of the commons according to the payoffs incentives described earlier. In the second stage (rounds 11–20) a new institution is introduced in the form of a regulation aimed at improving social earnings. Different forms of regulation were tested by varying the severity of the regulation (high vs low), and the enforcement mechanism (exogenous vs endogenous).

For the second round we explicitly told the players (a) that if each player chose to extract one unit the group would get the maximum earnings possible, and (b) that to enforce such a rule we would introduce a regulation by randomly choosing a player and applying a fine on the inspected player based on her deviation from the socially optimum solution. We varied the size of the fine, and the mechanism for introducing the regulation by allowing some groups to vote for the external regulation. We compared these treatments with the baseline where no new rule is introduced during the second stage. The details of how the stages were conducted follow.

Pre-game stage (instructions and practice rounds)

Each of the experiments begin by reading the instructions to the group of five players, and handing the experimental forms to each of them (see appendices): GAME CARDS (yellow) where they make their choice in each round, DECISIONS RECORDS SHEET (green) for their own calculations and record-keeping, and the PAYOFF TABLE (blue). Once all questions from participants are clarified, the experimenter conducts one demonstration round and one more practice rounds. After answering the questions, stage 1 begins.

Stage 1 (rounds 1–10)

In stage 1 of the experiment each player must decide privately her/his level of individual extraction from the commons, and write this down in a yellow round card; the same information is also recorded in the blue records sheet. The monitor collects the five cards, adds the total extraction for the group which he writes in the monitor's record sheet, and announces the total publicly. Each player must write the group's total, and by subtracting her/his individual extraction s/he is able to calculate the payoffs for that round. S/he writes her/his total gains for the round and the experiment proceeds to the next round by filling a new card.

Stage 2 (new rule, rounds 11–20)

During stage 2 we introduce the different forms of regulation of the use of the commons by varying the severity (high, low) of the regulation and the type (exogenous, endogenous) of enforcement. The stage will typically begin by an announcement from the monitor describing a new regulation to be enforced in each round aimed at producing the socially optimal point where the group's earnings are maximized, that is $X_i = 1$.

For all regulation treatments, the monitor began reading the instructions in the second stage by saying explicitly that by then all players would have noticed that the group can earn the maximum of points if all players chose to extract one unit only and that the new rule was aimed at producing such an outcome. Then the monitor continued with the details on how each different rule was to be administered. A portion of the groups were told that through a one-time vote they could decide by a simple majority (three or more votes)

to have such regulation applied to them from then on. For a portion of the groups the regulation was mandatory.

There were also control groups for a baseline treatment with no change in the rules for the second stage and where no mention was made of a socially desirable set of individual choices. Such baseline groups were not allowed to have any communication either.

The regulations were administered in the following way. After all players had calculated their earnings for a specific round, a player was chosen randomly from the group, followed by an inspection by the monitor to verify compliance with the new rule. If the majority of the group voted against the application of the regulation, the monitor would conduct the remaining rounds in the second stage in the same manner as in the first stage. However, the monitor would have told these groups about the rule to achieve the social optimum through each player choosing one unit of extraction.

For groups that had voted in favour of external regulation, or that had faced the same but mandatory rule, the inspection worked as follows. If there was compliance, the monitor would continue the experiment to the next round. If the inspected player chose more than one unit of extraction, the monitor would subtract from her total earnings the penalty p times the number of units in non-compliance. Once the player was selected publicly, the rest of the procedure was conducted in private between the monitor and the player, that is, the rest of the players in that round did not know what was actually chosen by the inspected player, nor her level of compliance or the penalty applied. Then the experiment proceeded to the next round until stage 2 was completed.

Exit stage (calculate earnings, fill out survey)

After the completion of all stage 2 rounds, the monitors calculated the total earnings for each player by adding the column of the earnings from the round and subtracting the cases where a fine was imposed. While the monitor made the calculations, the players responded to the exit survey anonymously and in private. Then payments were made privately in cash to each player, on the return of the filled out survey.

Community workshops

After all experiments had been conducted in each of the villages, and the research team had taken a day off for data analysis, a community workshop was held with all participants and other interested villagers for the presentation of the results, discussions about their plausible explanations, and the relation of the experiments with the institutional context of the village regarding the use of the commons and their regulation. In other words, the workshop helped in the collection of analytical stories about the internal and external validity of the experiments, which could later be tested with the experimental and survey data.

Experimental treatments for stage 2 (regulation)

The detailed experiments and treatment variables are explained in Table 11.1, where:

- X_t, is the level of extraction from the commons in round t. The decision on extraction is made individually and privately by each person.
- R, is the introduction of the regulation with a probability p of inspection to a player, and a fine f in case of non-compliance.
- V, is a voting process where the group decides by a majority of votes whether or not to introduce regulation with probability and fines p, f announced by the experimenter before the vote.

For instance $(X\text{-}V\text{-}R)_{11}$ means that in round 11 for this group, each player individually decides the level of extraction (X_i) and her vote (V_i) to apply the regulation in that round. Once the decisions are collected and the votes counted, the monitor applies the regulation (R), if three or more voted 'yes'.

The sample of groups was distributed across treatments as listed in Table 11.2.

Results: behavioural consequences of external regulations

For presenting the experimental results and deriving some final insights from them, I present a sequence of results that offer some lessons in at least two categories. First, there are methodological lessons regarding the use of economic experiments to explore questions that involve behavioural and psychological elements with policy implications. Secondly, there are lessons regarding the design and effects of external regulations aimed at producing socially desired outcomes, but with incomplete enforcement.

According to the canonical model, the behavioural prediction for individuals maximizing their own material payoffs is that each chooses eight units of extraction for a group extraction of 40 units. Such prediction is contrasted with the socially optimum level of group extraction of five units.

Let us first look at the experimental results for the mandatory regulations under a strong and weak fine mechanism. Figure 11.1 shows the evolution over rounds for the average group extraction under a sub-set of treatments that allows one to appreciate the first set of results. For clarity the data for the baseline treatments and for the mandatory regulations with low and high fines are included. The data for the regulations subject to a group vote will be shown and discussed later. All dotted lines in the figure represent data for the student groups, while the solid lines reflect the data for the field experiments conducted with villagers. Larger triangles represent the high fine (XHR) data, while smaller triangles represent data for the lower fine.

Table 11.1 Experimental design, stages and treatment

Experimental treatments	*Stage* 1 (10 *rounds*)	*Stage* 2 (10 *rounds*)		
		New rule	(*Round* 11)	(*Rounds* 12–20)
(X) Baseline	$X_1, X_2, X_3, \ldots, X_{10}$	(No change – control)	X_{11}	$X_{12}, X_{13}, \ldots, X_{20}$
(XRL) Exogenous regulation (weak enforcement)	$X_1, X_2, X_3, \ldots, X_{10}$	Inspection of 1 player in each round: $p = 1/5, f = \$50$	$(X\text{-}R)_{11}$	$(X\text{-}R)_{12}, (X\text{-}R)_{13}, \ldots, (X\text{-}R)_{20}$
(XRH) Exogenous regulation (strong enforcement)	$X_1, X_2, X_3, \ldots, X_{10}$	Inspection of 1 player in each round: $p = 1/5, f = \$175$	$(X\text{-}R)_{11}$	$(X\text{-}R)_{12}, (X\text{-}R)_{13}, \ldots, (X\text{-}R)_{20}$
(VXRL) Endogenous (voted) regulation (weak enforcement)	$X_1, X_2, X_3, \ldots, X_{10}$	Inspection of 1 player in each round: $p = 1/5, f = \$50$	$(V\text{-}X\text{-}R)_{11}$	$(X\text{-}R)_{12}, (X\text{-}R)_{13}, \ldots, (X\text{-}R)_{20}$
(VXRH) Endogenous (voted) regulation (strong enforcement)	$X_1, X_2, X_3, \ldots, X_{10}$	Inspection of 1 player in each round: $p = 1/5, f = \$175$	$(V\text{-}X\text{-}R)_{11}$	$(X\text{-}R)_{12}, (X\text{-}R)_{13}, \ldots, (X\text{-}R)_{20}$

Table 11.2 Sample size by treatments

Experiment (treatment)	*Sample size by treatment*			
	Villagers		*Students*	
	No. groups	*No. people*	*No. groups*	*No. people*
(X) Baseline	8	40	3	15
(XRL) Exogenous regulation (weak enforcement)	20	100	5	25
(XRH) Exogenous regulation (strong enforcement)	12	60	4	20
(VXRL) Endogenous regulation (weak enforcement)	12	60	5	25
(VXRH) Endogenous regulation (strong enforcement)	12	60	4	20
Sub-totals	64	320	21	105

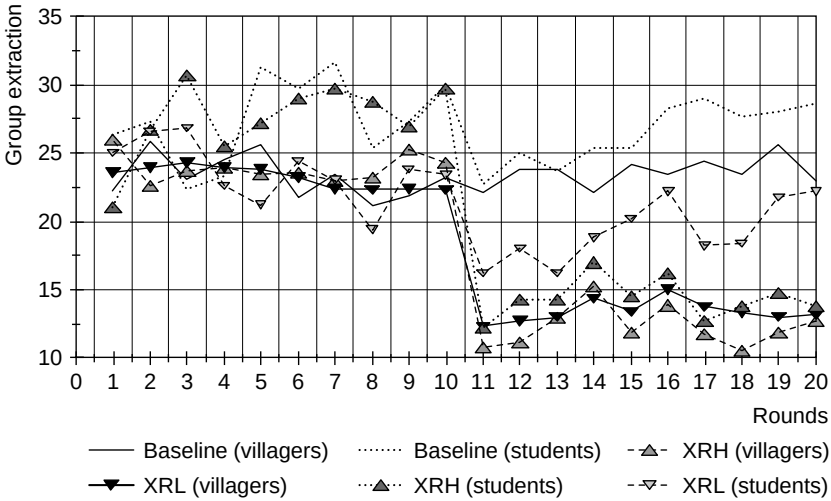


Figure 11.1 Average group extraction over time by treatments and subject pools

Result 1 Under a baseline treatment, players neither act accordingly to the canonical *homo economicus* model, nor achieve the socially optimum condition. Students and villagers do show a similar pattern of behaviour under the baseline treatment, with a slight but statistically significant shift towards more cooperative behaviour among the villagers.

In Figure 11.1 the lines with no triangles represent the average group extraction for the baseline treatments. Throughout the rounds we can observe a slightly lower level of group extraction for the villagers (a two-sample Wilcoxon rank-sum Mann–Whitney test yields a p-value = 0.000 when testing individual decisions data, and 0.0001 for group-level data[6]). Given that the socially optimum level of group earnings in a round, according to the payoffs table, is $(758 \times 5) = 3{,}790$ points, and that on average the students groups under the baseline treatments earned 2,498.3 and villagers groups earned on average 2,727.7, we find that students achieved a level of social efficiency of about 66 per cent while villagers earned about 72 per cent, compared to the 42 per cent of efficiency predicted at the Nash equilibrium of maximum feasible extraction. Further, while about 25 per cent of the student decisions fell exactly within the predicted eight units of extraction, only 14 per cent of these Nash choices were made by villagers. Below I outline plausible explanations for this result, but in general it is more likely that the more experience villagers have of such situations the more it will induce them to refrain from overextracting. In fact we can observe a slight but steady trend towards higher extraction by students during the last rounds and a flatter trend for villagers.

Result 2 There are positive effects of a partially enforced regulation based on a penalty on overextraction, but the effect of the penalty level does not seem to fit the expected cost prediction (based on the expected utility model). Players seem to cooperate 'too much' under a low level of penalty, but they also seem to free-ride 'too much' under a higher penalty.

Continuing with Figure 11.1, we move now to the case of the externally imposed regulation with two variations. Under variation one, the weaker regulation (XRL), with a lower fine of $f = 50$ points per unit above one unit of extraction, and a probability of inspection of 1/5 in each round, should shift the material payoff maximizers from eight to six units, according to the model of expected costs described earlier. Under variation two, the stronger regulation (XRH) with exactly the same conditions except for a fine of $f = 175$ points would create an expected cost high enough to induce the players to choose the socially optimum condition of one unit of extraction, namely the same as their Nash strategy.

The results, presented in the graphs and descriptive statistics, suggest that such regulations do induce players to reduce their average level of extraction, particularly for the early rounds of the second stage. Compared to baseline data in the second stage, all treatments under the weak or heavier regulations show a significant reduction in extraction. However, a few interesting phenomena deserve attention.

First, in both cases, low and high fine, the average villager shows a lower level of extraction during the second stage of the sessions. With the low

fine (XRL), on average villagers chose about 1.173 units below the students, while in the case of the high fine (XRH) the difference is only 0.537 units although still significant. The 2-sample Wilcoxon rank-sum Mann–Whitney tests respectively yield p-values of 0.0000, and 0.0002 for individual-level data and 0.0000 and 0.0086 for group-level data. It is particularly interesting to note the early compliance by students under the low fine, and how their compliance decreases over time, almost reaching the level of extraction they were showing before the second stage of the game.

It is also important to observe the small difference between the low and the high levels of fines, as compared with the prediction from the model that those under the high fine XRH, if choosing their best response, should play at the same level of the social optimum ($x = 1$ unit) while those under the low fine should choose six units of extraction. However, the difference with different levels of penalty is very small, and suggests that the players do not calculate the costs of regulations in the same manner. In fact the difference is substantially smaller for villagers than for students. On average over the rounds, villagers chose 0.34 units less under the high fine while students chose 0.981 units less under the same high fine. Recall that the predicted difference between the weak and the strong-fine Nash strategy for players is five units, that is, 25 units of difference for the group extraction, between the two symmetric Nash equilibria. Apparently players under the low-fine XRL seem to be cooperating more than they should, while those under the high penalty (XRH) seem to be attempting to capture rents by taking the chances of 4/5 of not being inspected.

For both students and villagers the differences were statistically significant and in the expected direction, but the size of the difference, and the fact that it is smaller for the villagers, does suggest that our participants were following rules of thumb, suggested by the external authority – the experimenter in this case – in addition to their calculation of the benefits and costs of being inspected and regulated (see Cardenas, 2004a, for a further discussion of this phenomenon).

Result 3 When allowed to vote (VXRL, VXRH), the majority of villagers oppose the externally imposed regulations while the majority of students vote in favour of introducing them. Further, for the groups in which the majority favoured regulation, the average behaviour was not different from the mandatory system of penalties (XRL, XRH).

The next set of treatments to be presented and discussed involve cases where the five players were asked to vote on whether they would like to face the regulation described for treatments XRL and XRH above. If three or more vote in favour, the regulation would be implemented in the same manner as in treatments XRL and XRH, but if three or more vote against such regulation,

the session would continue in exactly the same manner as in the first stage for the remaining 10 rounds.

A first prediction is that all players would vote in favour of the regulation, even in the case of selfish and risk-taking players. The rationale is simple. Players should expect others to follow the new rule, and to take the risk of not being inspected and choose a higher level of extraction. The returns from such a decision would be substantial, according to the payoffs table. Notice that voting and enforcing these regulations are costless for all five players, and therefore the expected cost of the regulation is equivalent to the expected cost in the mandatory case, while the dominant strategy continues to be greater individual extraction up to the symmetric Nash equilibrium.

Let us first look at voting behaviour at the group and individual level. Table 11.3 summarizes the number of groups that passed and failed to pass the votes among the villagers and among the students, for the two levels of penalties, high (VXRH) and low (VXRL).

Clearly, students were more prone to ask for these regulations than the villagers. Among the villagers, in 67 per cent of the 24 groups of villagers, the majority voted against the regulations at the group level and 64 per cent voted against regulations at the individual level. Among the students, however, 67 per cent favoured regulation at the group level and 71 per cent favoured regulations at the individual level. Also, there are some interesting differences across levels of penalties. Villagers were especially opposed to high fines while students were more inclined to vote for them. These results were confirmed at the community workshops we held days later in these villages, where high fines were regarded as too high and unnecessary.

If we look at individual voting, by subject pool and treatment (see Table 11.4) we observe how for villagers (the first two rows) only two groups obtained a total of four votes and none had all five votes in favour of regulation. In contrast, six out of the nine sessions for students got four or more votes.

Nevertheless, this difference in aversion to external regulations contrasts with the opposite intentions in choices. We have already observed in

Table 11.3 Distribution of voting by treatment and by subject pool

	% Groups that passed the rule						**% Individual votes**					
	Villagers			*Students*			*Villagers*			*Students*		
	No (%)	*Passed* (%)	*Total*	*No* (%)	*Passed* (%)	*Total*	*No* (%)	*Yes* (%)	*Total*	*No* (%)	*Yes* (%)	*Total*
VXRH	83	17	12	25	75	4	70	30	60	20	80	20
VXRL	50	50	12	40	60	5	58	42	60	36	64	25
Total	67	33	24	33	67	9	64	36	120	29	71	45

Table 11.4 Distribution of groups by total of votes in the group

	Frequency of groups by individual votes						
	0	*1*	*2*	*3*	*4*	*5*	*Total groups*
VXRH-villagers	2	4	4	2			12
VXRL-villagers	3	1	2	4	2		12
VXRH-students			1		1	2	4
VXRL-students		1	1		2	1	5
Total	5	6	8	6	5	3	33

the previous results that not only did the villagers show a slightly more cooperative behaviour for the baseline, but that their compliance with the suggested regulation was on average higher than among the students, despite their lower votes in favour of such enforcement. In the next result I will explore the individual decisions (extraction), comparing groups that did not pass the regulation and those that did with three or more votes.

Result 4 Players voting against the external regulation showed a significant willingness to cooperate under the suggested new rule. However, negative reciprocity and the absence of any other coordination mechanism generated a partial erosion of these group-oriented motivations, although still producing socially superior results than the first stage.

Figures 11.2 and 11.3 below describe the evolution of group extraction over rounds for all the sessions in which voting took place (VXRL and VXRH). Figure 11.2 describes the behaviour of those groups in which the majority voted against the regulation and therefore continued after round 11 under the same conditions as during the first stage. In other words, material incentives would be equivalent to the baseline treatment. Figure 11.3 illustrates the case of those groups that approved the regulation and therefore faced a penalty that had a probability of enforcement.

The behavioural effects of the regulations warrant some discussion. First of all, as in the previous analysis it seems that the effect of lower or higher fines does not create very large differences in behaviour. Whether the fine is of 50 points or 175 points, the average group effect is similar.

Two interesting phenomena, however, seem to occur where regulations are subject to a group vote. First, all players, regardless of whether they approve or reject the regulation, showed a significant decrease in extraction by round 11 and for the subsequent few rounds. Notice that before the decision is made in round 11, all groups were told of the purpose of the new rule. Second, by the end of round 11 they learnt if the regulation was to be implemented for

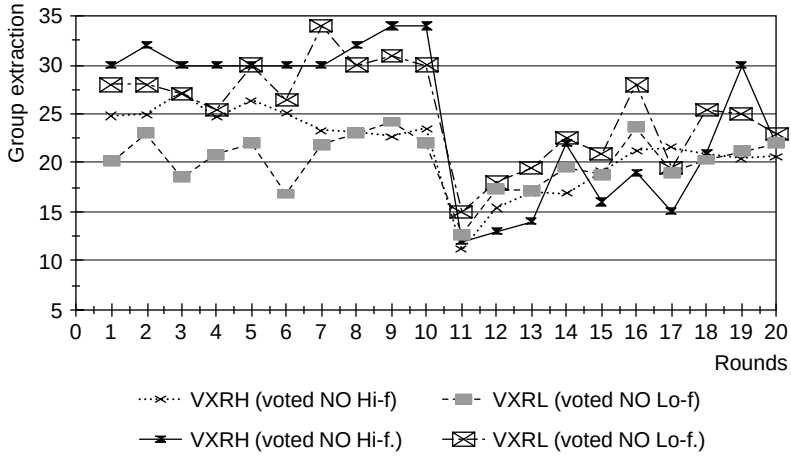


Figure 11.2 Average group extraction for groups that did NOT pass the proposed regulations

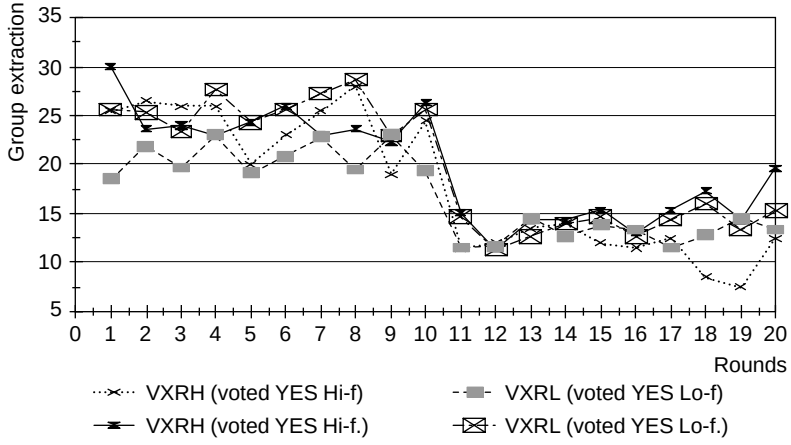


Figure 11.3 Average group extraction for groups that DID pass the proposed regulations

the rest of the session. The willingness of the participants (whether students or villagers) to cooperate, emerges as an effect of the norm 'proposed' by the external regulator, namely, the monitor who announced out loud to the group that the purpose of the new regulation being proposed was to achieve maximum earnings for the group.

Therefore, all groups witnessed a strong shift towards the group-maximizing outcome regardless of the voting. Further, the average player among the groups that did not pass the rule still showed a willingness to follow the group-maximizing strategy. However, there was an erosion of cooperation among these groups that did not pass the regulation after the first round of attempts at a non-binding low level of extraction, while the groups under the approved regulation maintained the lower levels of extraction over the rounds. The more plausible explanation for this behaviour is the triggering of negative reciprocity by players who witnessed other players to be extracting more than was suggested, and therefore using their own increase in extraction either as a punishing device or at least to avoid the sucker's payoff, which has been consistently found in the literature. Nevertheless, all groups that did not pass the vote ended the second stage under lower levels of extraction than at the end of their first stage.

Now comparing the subject pools for cases that did not approve the regulation, we find no significant differences, but for the approved sessions. Villagers on average, at the end of the second stage, showed a slightly lower level of group extraction.[7] Recall that villagers were less prone than students to vote in favour of such regulations and these regulations were approved in only a few villages.

Enriching experiments with other field techniques

The additional fieldwork tools can help explain some of the results and puzzles just presented, particularly with respect to bias against regulations and yet a willingness to reduce extraction in the experiment. The workshops and follow-up field interviews we conducted after the end of the experiments in each village allowed us to discuss the low voting in favour of a rule that would clearly increase players' earnings. It also helped us to explore why players on average reduced their extraction even where the group voted to continue playing under the same rules as in the first stage. The analogies with the actual context of rules and presence of regulating agencies were clear. By exploring the differences across the three villages we could deepen this analysis.

In general, a majority of village groups were against the regulations, but there were some differences across sites that were discussed during the community workshops in which we used rapid rural appraisal techniques. In general, in only eight out of the 24 village groups did the majority vote in favour of the regulation, while among the students the ratio was six out of nine. In terms of individual votes, 32 students out of 45 voted in favour, while only 43 of the 120 villagers did so.

Consider in more detail the three villages where the voting took place. In Neusa, only one out of eight groups had a majority vote, in Sanquianga

four out of eight groups voted in favour, and in LaVega three of the five groups voted in favour of the different rules. During the workshops we used some tools to study, in each case, the relations of external actors with the community of resource-users, including state and non-government organizations. Neusa represented the most conflict-ridden case of a close but negative relation with the *Corporacion Autonoma Regional*, the environmental agency in charge of monitoring and controlling the use and extraction of resources. For instance, in Neusa the regulatory agency had hired a private security company to monitor the fishermen's catch every day. In LaVega the regulatory agency was basically absent as far as the villagers were concerned and was replaced by a non-government organization, the regional federation of coffee growers, whose role was to provide support and technical assistance rather than to regulate. In Sanquianga, again, the regional authority was absent; instead there was the National Parks System (Ministry of Environment) with a closer but more positive relation with the community of fishermen and clam gatherers, given its lack of resources, funding and weak political clout for enforcing rules through monitoring and sanction.

5 Discussion and conclusions

There is surprisingly little work on the behavioural response of individuals to externally imposed regulations, such as those imposed by state agencies, that issue rules that are aimed at maximizing social welfare but that can only partially enforce compliance. Also, the experimental literature still needs to open the debate when finding behavioural differences across subject pools. Such a debate could enrich the understanding of phenomena we find in the economic laboratory.

The experimental work presented in this chapter, conducted in the laboratory with students and in the field with villagers that have joint access to a natural resource, is aimed at filling these gaps to some extent. By comparing these two different subject pools we observe similar behaviour, but with a statistically significant difference in attitudes towards externally imposed regulations. The villagers rejected such regulations more often than students when allowed to vote for their implementation, but were more willing to cooperate under a non-binding setting.

The results leave some puzzles unsolved and in fact open new questions for research. How rules affect behaviour and therefore outcomes cannot be explained easily through the basic game-theoretic model of maximizing material incentives, or the use of expected utility theory. At least in our experimental results, individuals respond to more factors than just the expected cost of regulations. Our field evidence and our discussions and workshops

with the field participants, also indicate that there is a bias against the imposition of regulations that can clearly help each group member increase her/his earnings. Our field discussions also suggest that the effect of state regulators can often be negative in terms of inducing a self-governed solution to the commons dilemma, particularly where regulation by the state is costly and difficult.

As one of the women – a gatherer of molluscs in the Sanquianga mangrove forest region – said at the end of her session in which we allowed the group to vote for the regulation, *'I didn't want the rule or the penalty, but I do want to pick the large ones only'*, referring to how by picking only the large clams the resource could naturally reproduce and maintain a sustainable yield. In her case, as in several others, the second stage showed a decrease in extraction even though the majority voted against the regulation.

The differences in behaviour and outcomes between students and villagers can also open new methodological implications for experimental research. I believe these two are complementary and not necessarily contradictory. It seems that student behaviour is slightly closer to the game-theoretical predictions. If this is indeed the case, we need to discuss the implications further. Often experimental studies are designed to illustrate policy design. If conducted only with students, without replication in the field with subjects that are familiar with the problem in practical terms, the studies can miss relevant information. Further, some experiments can be used to enlighten policy designs which are more likely to be applied to people in the field than to students.

In this sense, the field can help explain the variations across and within groups observed in many of these experiments. Having the villagers participate in the research, by discussing with the research team the preliminary data from the experiments, helped us clarify some of the key puzzles. By comparing the villages and the particular context in which the participants face external regulations from state agencies and other actors, we pose plausible differences in the fraction of players voting in favour of experimental regulations. I discuss elsewhere (Cardenas, 2004b) how bringing the laboratory to the field can enrich our experimental work, as people in experiments bring with them information in the form of clues or rules of thumb based on their own experience in similar situations. In the field experiments, the context of our villagers is definitely providing some hints to our participants during the game, as they clearly explained to us in our interviews and workshops. The context and prior experience of students is different from that of the villagers, leading to behavioural differences in the experiments. This brings us to the last issue, namely the rich information that can emerge when we combine experimental work in the field with that in the university laboratory. This would make the policy conclusions derived from experiments broader and more relevant than if only students are used as subject pools, as has been typical of behavioural economics to date.

Appendix

Payoffs table experimental design

		MY LEVEL OF EXTRACTION								
		1	2	3	4	5	6	7	8	Their average extraction
THEIR EXTRACTION LEVEL	4	758	790	818	840	858	870	878	880	1
	5	738	770	798	820	838	850	858	860	1
	6	718	750	778	800	818	830	838	840	2
	7	698	730	758	780	798	810	818	820	2
	8	678	710	738	760	778	790	798	800	2
	9	658	690	718	740	758	770	778	780	2
	10	638	670	698	720	738	750	758	760	3
	11	618	650	678	700	718	730	738	740	3
	12	598	630	658	680	698	710	718	720	3
	13	578	610	638	660	678	690	698	700	3
	14	558	590	618	640	658	670	678	680	4
	15	538	570	598	620	638	650	658	660	4
	16	518	550	578	600	618	630	638	640	4
	17	498	530	558	580	598	610	618	620	4
	18	478	510	538	560	578	590	598	600	5
	19	458	490	518	540	558	570	578	580	5
	20	438	470	498	520	538	550	558	560	5
	21	418	450	478	500	518	530	538	540	5
	22	398	430	458	480	498	510	518	520	6
	23	378	410	438	460	478	490	498	500	6
	24	358	390	418	440	458	470	478	480	6
	25	338	370	398	420	438	450	458	460	6
	26	318	350	378	400	418	430	438	440	7
	27	298	330	358	380	398	410	418	420	7
	28	278	310	338	360	378	390	398	400	7
	29	258	290	318	340	358	370	378	380	7
	30	238	270	298	320	338	350	358	360	8
	31	218	250	278	300	318	330	338	340	8
	32	198	230	258	280	298	310	318	320	8

Experiment instructions (English translation)

These instructions were originally written in Spanish and translated from the final version used in the field work. The instructions were read to the participants from the script below by the same person during all sessions. The participants could interrupt and ask questions at any time.

Whenever the following type of text and font e.g. [. . .MONITOR: distribute **PAYOFFS TABLE** to participants. . .] is found below, it refers to specific instructions to the monitor at that specific point; when in *italics*, these are notes added to clarify issues to the reader. Neither of these were read to participants. Where the word 'poster' appears, it refers to a set of posters we printed in very large format with the payoffs table, forms, and the

three examples described in the instructions. These posters were hung on a wall near the participants' desks where the eight people could see them easily.

COMMUNITY RESOURCES GAME (Instructions)

Greetings...
We want to thank every one here for attending the call, and specially thank the field practitioner _____ (*name of the contact person in that community*), and _____ (*local organization that helped in the logistics*) who made this possible. We will spend about two hours between explaining the exercise, playing it and finishing with a short survey at the exit. So, let us get started.

The following exercise is a different and entertaining way of participating actively in a project about the economic decisions of individuals. Besides participating in the exercise, and being able to earn some prizes and some cash, you will participate in a community workshop in two days to discuss the exercise and other matters about natural resources. During the day of the workshop we will give you what you earn during the game. The funds to cover these expenditures have been donated by various international organizations and the University.

1 Introduction

This exercise attempts to recreate a situation where a group of families must make decisions about how to use the resources of, for instance, a forest, a water source, a mangrove, a fishery, or any other case where communities use a natural resource. In the case of this community _____ (name of the specific village), an example would be the use of firewood or logging in the _____ (name of an actual local commons area in that village) zone. You have been selected to participate in a group of five people among those that signed up for playing. The game in which you will participate now is different from the ones others have already played in this community, thus, the comments that you may have heard from others do not apply necessarily to this game. You will play for several rounds equivalent, for instance, to years or harvest seasons. At the end of the game you will be able to earn some prizes in kind and cash. The cash prizes will depend on the quantity of points that you accumulate after several rounds.

1 The PAYOFFS TABLE

To be able to play you will receive a **PAYOFFS TABLE** equal to the one shown in the poster. [...MONITOR: show **PAYOFFS TABLE** in poster and distribute **PAYOFFS TABLE** to participants...]

This table contains all the information that you need to make your decision in each round of the game. The numbers that are inside the table correspond to points (or pesos) that you would earn in each round. The only thing that each of you has to decide in each round is the LEVEL OF EXTRACTION that you want to allocate extracting resources(in the columns from 1 to 8).

To play in each round you must write your decision number between 1 and 8 in a yellow GAME CARD like the one I am about to show you. [. . .MONITOR: show **yellow GAME CARDS** and show in the poster. . .] It is very important that we keep in mind that the decisions are absolutely individual, that is, that the numbers we write in the game card are private and that we do not have to show them to the rest of members of the group if we do not want to. The monitor will collect the 5 cards from all participants, and will add the total units of extraction that the group decided to allocate. When the monitor announces the group total, each of you will be able to calculate the points that you earned in the round. Let us explain this with an example.

In this game we assume that each player extract as maximum of 8 units of a resource like firewood or logs. In reality this number could be larger or smaller but for purposes of our game we will assume 8 as maximum. In the PAYOFFS TABLE this corresponds to the columns from 1 to 8. Each of you must decide from 1 to 8 in each round. But to be able to know how many points you earned, you need to know the decisions that the rest in the group made. That is why the monitor will announce the total for the group in each round. For instance, if you decide to extract 2 units and the rest of the group together, add to 20 units, you would gain ____ points. Let us look at two other examples in the poster.

[. . .MONITOR: show poster with the **THREE EXAMPLES**…]

Let us look how the game works in each round.

2 The DECISIONS FORM

To play each participant will receive one green DECISIONS FORM like the one shown in the poster in the wall. We will explain how to use this sheet. [. . .MONITOR: show the **DECISIONS FORM** in the poster and distribute the **DECISIONS FORMS**. . .]

With the same examples, let us see how to use this DECISIONS FORM. Suppose that you decided to play 5 units in this round. In the yellow GAME CARD you should write 5. Also you must write this number in the first column A of the decisions form. The monitor will collect the 5 yellow cards and will add the total of the group. Suppose that the total added 26 units. Thus, we write 26 in the column B of the decisions form. [. . .MONITOR: In the poster, write the same example numbers in the respective cells. . .]

To calculate the third column (C), we subtract from the group total, MY DECISION and then we obtain THEIR LEVEL OF EXTRACTION which we write in column C. In our example, $26 - 5 = 21$. If we look at the PAYOFFS TABLE, when MY EXTRACTION are 5 and THEIR EXTRACTION are 21, I earn ____ points. I write then this number in the column D of the DECISIONS FORM.

It is very important to clarify that nobody, except for the monitor, will be able to know the number that each of you decides in each round. The only thing announced in public is the group total, without knowing how each participant in your group played. Let us repeat the steps with a new example. [. . .MONITOR: Repeat with the other two examples, writing the numbers in the posters hanging in the wall. . .]

It is important to repeat that your game decisions and earnings information are private. Nobody in your group or outside of it will be able to know how many points you earned or your decisions during rounds. We hope these examples help you understand how the game works, and how to make your decisions to allocate your UNITS OF EXTRACTION in each round of the game. **If at this moment you have any question**

about how to earn points in the game, please raise your hand and let us know. [. . .MONITOR: pause to resolve questions. . .]

It is very important that while we explain the rules of the game you do not engage in conversations with other people in your group. If there are no further questions about the game, then we will assign the numbers for the players and the rest of forms needed to play.

2 Preparing for playing

Now write down your player number in the green DECISIONS FORM. Write also the place _____ and the current date and time __/__/__, __:__am/pm. In the following poster we summarize for you the steps to follow to play in each round. Please raise your hand if you have a question. [MONITOR: Read the steps to them from the poster]

Before we start, and once all players have understood the game completely, the monitor will announce one additional rule for this group. To start the first round of the game we will organize the seats and desks in a circle where each of you face outwards. The monitor will collect your yellow game cards in each round. Finally, to get ready to play the game, please let us know if you have difficulties reading or writing numbers. If so, one of the monitors will sit next to you and assist you with these. Also, please keep in mind that from now on there should be no conversation nor should statements be made by you during the game, unless you are allowed to. We will first have a few rounds of practice that will NOT count toward your real earnings, they are just for practicing the game.

Notes

1 The experimental design and treatments were inspired by previous results (Cardenas *et al*., 2000) where similar experiments in the field produced the intriguing result of a crowding-out of group-oriented preferences.
2 An analysis of the experimental data across sites goes beyond the scope of this chapter.
3 This chapter does not deal with other objective functions for the planner or regulator, such as private or political rent-seeking behaviour.
4 See e.g. Falkinger, Fehr, Gachter and Winter-Ebmer (2000); Fehr and Gachter (2000); Sefton, Shupp and Walker, (2000); Ostrom, Gardner and Walker (1994); Ostrom, Walker and Gardner (1992); Bohnet, Frey and Huck (2001); and Cason and Kahn (1999).
5 Evidence from other experiments conducted in the field is growing (Henrich, 2000; Henrich *et al*., 2001; Ensminger, 1999) but so far such experiments have not tested regulations aimed at correcting individual behaviour.
6 Individual-level tests use individual choices in each round as observations; group-level tests are conducted with group totals in each round serving as observations.
7 Only in the case of high penalty that passed the vote (VXRH (voted YES hi-f) is there a significant difference with villagers showing a lower level of group extraction.

References

Andreoni, J., W. T. Harbaugh and L. Vesterlund (2003) 'The Carrot or the Stick: Rewards, Punishments and Cooperation', *American Economic Review*, vol. 93(3), pp. 893–902.
Baland, J.-M. and J.-P. Platteau (1996) *Halting Degradation of Natural Resources: Is There a Role For Rural Communities?* (Oxford and New York: Oxford University Press).

Berkes, F. (ed.) (1989) *Common Property Resources: Ecology and Community-Based Sustainable Development* (London: Belhaven Press).
Bohnet, I., B.S. Frey and S. Huck (2001) 'More Order with Less Law: On Contract Enforcement, Trust, and Crowding', *American Political Science Review*, vol. 95(1), pp. 131–44.
Camerer, C. and E. Fehr (2004) 'Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists,' in J. Heinrich *et al.*, pp. 55–95.
Cardenas, J.C. (2004a) 'Norms from Outside and from Inside: An Experimental Analysis on the Governance of Local Ecosystems', *Forest Policy and Economics*, vol. 6, pp. 229–41.
Cardenas, J.C. (2004b) 'Bringing the Lab to the Field: More Than Changing Subjects' Paper presented at the International Meeting of the Economic Science Association, Pittsburgh, June.
Cardenas, J.C., J.K. Stranlund and C.E. Willis (2000) 'Local Environmental Control and Institutional Crowding-out', *World Development*, vol. 28(10), pp. 1719–33.
Carpenter, J. (forthcoming) 'Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods', *Games and Economic Behavior*.
Carpenter, J. and P. Mathews (2004) 'Social Reciprocity', mimeo, Dept of Economics, Middlebury College.
Cason, T.N. and F.U. Kahn (1999) 'A Laboratory Study of Voluntary Public Goods Provision with Imperfect Monitoring and Communication', *Journal of Development Economics*, vol. 58, pp. 533–52.
Ensminger, J. (2000) 'Experimental Economics in the Bush: Why Institutions Matter', in C. Menard (ed.), *Institutions, Contracts and Organizations: Perspectives from New Institutional Economics* (London, Edward Elgar), pp. 158–71.
Falkinger, J., E. Fehr, S. Gachter and R. Winter-Ebmer (2000) 'A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence', *American Economic Review*, vol. 90(1), pp. 247–64.
Fehr, E. and D. Schmidt (1999) 'A Theory of Fairness, Competition, and Cooperation', *Quarterly Journal of Economics*, vol. 114, pp. 817–51.
Fehr, E. and S. Gachter (2000) 'Do Incentive Contracts Crowd Out Voluntary Cooperation?', Institute for Empirical Research in Economics, University of Zürich, Working Paper no. 34.
Gintis, H. (2000) 'Beyond Homo Economicus: Evidence from Experimental Economics', *Ecological Economics*, vol. 35, pp. 311–22.
Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis and R. McElreath (2001) 'Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies', *American Economic Review,* vol. 91, pp. 73–8.
Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis (2004) *Foundations of Human Sociality; Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (Oxford: Oxford University Press).
Kahneman, D. and A. Tversky (eds) (2000) *Choices, Values and Frames* (Cambridge: Cambridge University Press for Russell Sage Foundation).
Loewenstein, G. (1999) 'Experimental Economics from the Vantage Point of Behavioural Economics', *Economic Journal*, vol. 109, pp. F25–F34.
Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge and New York: Cambridge University Press).
Ostrom, E. *et al*. (eds) (2002) *The Drama of the Commons*, Committee on the Human Dimensions of Global Change (Washington DC: National Academy Press).

Ostrom, E., J. Walker and R. Gardner (1992) 'Covenants With and Without a Sword: Self-governance Is Possible', *American Political Science Review*, vol. 86(2), pp. 404–17.

Ostrom, E., R. Gardner and J. Walker (1994) *Rules, Games and Common-Pool Resource* (Ann Arbor: University of Michigan Press).

Rabin, M. (1998) 'Psychology and Economics', *Journal of Economic Literature*, vol. 36, pp. 11–46.

Rabin, M. (2002) 'A Perspective on Psychology and Economics. Alfred Marshall Lecture', *European Economic Review*, vol. 46, pp. 657–85.

Sefton, M., R. Shupp and J. Walker (2002) 'The Effect of Rewards and Sanctions in Provision of Public Goods', Working Paper no. 2002–02, Centre for Decision Research and Experimental Economics, University of Nottingham, UK.