Experimental Economics

Method and Applications

Over the past two decades, experimental economics has moved from a fringe activity to become a standard tool for empirical research. With experimental economics now regarded as part of the basic tool-kit for applied economics, this book demonstrates how controlled experiments can be useful in providing evidence relevant to economic research. Professors Jacquemet and L'Haridon take the standard model in applied econometrics as a basis for the methodology of controlled experiments. Methodological discussions are illustrated with standard experimental results. This book provides future experimental practitioners with the means to construct experiments that fit their research question, and newcomers with an understanding of the strengths and weaknesses of controlled experiments. Graduate students and academic researchers working in the field of experimental economics will be able to learn how to undertake, understand and criticise empirical research based on lab experiments, and refer to specific experiments, results or designs completed with case study applications.

Nicolas Jacquemet is a full professor at University Paris-1 Panthèon Sorbonne and the Paris School of Economics. His research combines experimental methods and econometrics to study discrimination, the effect of personality traits on economic behaviour, the role of social pre-involvement in strategic behaviour and experimental game theory. His research has been published in *Econometrica, Management Science, Games and Economic Behavior, Journal of Environmental Economics and Management, Journal of Health Economics* and *Journal of Economic Psychology.*

Olivier L'Haridon is a full professor at the University of Rennes 1. His research combines experimental methods and decision theory, applied in the study of individual decision-making as affected by uncertainty. His work has been published in *American Economic Review, Management Science, Journal of Risk and Uncertainty, Theory and Decision, Experimental Economics, Journal of Health Economics* and *Journal of Economic Psychology.*

Experimental Economics

Method and Applications

NICOLAS JACQUEMET
University Paris 1 Panthèon-Sorbonne and Paris School of Economics, France

OLIVIER L'HARIDON
Université de Rennes I, France

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107060272
DOI: 10.1017/9781107446786

First published 2018

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

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

Library of Congress Cataloging-in-Publication Data
Names: Jacquemet, Nicolas, author. | L'Haridon, Olivier, author.
Title: Experimental economics method and applications / Nicolas Jacquemet, Paris School of Economics, Olivier L'Haridon, Université de Rennes I, France.
Description: Cambridge, United Kingdom ; New York, NY, USA : Cambridge University Press, [2018] | Includes bibliographical references and index.
Identifiers: LCCN 2018007008 | ISBN 9781107060272
Subjects: LCSH: Experimental economics.
Classification: LCC HB131 .J33 2018 | DDC 330.072/4–dc23
LC record available at https://lccn.loc.gov/2018007008

ISBN 978-1-107-06027-2 Hardback
ISBN 978-1-107-62977-6 Paperback

Contents

Figures

Tables

Illustrations

Focuses

Abbreviations and Symbols

Abbreviations

AD	Aggregate Demand
ATE	Average Treatment Effect
ATT	Average Treatment on the Treated
BART	Balloon Risk Analogue Task
BDM	Becker-De Groot-Marschak
BMI	Body Max Index
CADI	Constant Absolute Decreasing Impatience
CDF	Cumulative Distribution Function
CE	Certainty Equivalence
CHM	Cognitive-Hierarchy Model
CRDI	Constant Relative Decreasing Impatience
CRRA	Constant Relative Risk Aversion
DARA	Decreasing Absolute Risk Aversion
DA	Deferred Acceptance algorithm
DGP	Data Generating Process
DM	Dissonance Minimization
ECU	Experimental Currency Unit
FPRP	False Positive Report Probability
FR	Fully-revealing game
FTC	Federal Trade Commission
FW	Fixed wage
HSD	Honestly Significant Difference
IEC	Institutional Ethics Committee
IOS	Inclusion of the Other in the Self
IQR	Interquartile Range
IRB	Institutional Review Board
IV	Induced Value
LHS	Left-Hand Side
LSD	Least Significant Difference
MARS	Meta-Analysis Reporting Standards
MD	Mean absolute Deviation
MLE	Maximum Likelihood Estimator
MOOSE	Meta-analysis of Observational Studies in Epidemiology

MPCR	Marginal per Capita Return
MSE	Mean Squared Error
MT	Amazon's Mechanical Turk
MT	Mechanical Turk
MT	Western Educated, Industrialized, Rich, and Democratic
NR	Non-revealing game
OLS	Ordinary Least Squares
PEEM	Portable Extensions of Existing Models
PE	Probability Equivalence
PGG	Public Good Game
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PR	Piece-rate
Q-Q	Quantile-Quantile
QRE	Quantal-Response Equilibrium
RDU	Rank-Dependent Utility
RHS	Righ-Hand Side
RIS	Random Incentive System
UBG	Ultimatum Bargaining Game
VCM	Voluntary Contribution Mechanism
WEIRD	Western Educated, Industrialised, Rich, and Democratic
WTA	Willingness to Accept
WTP	Willingness to Pay
WVS	World Value Survey
WVS	World Values Survey

Symbols

$\bar{y}$	sample average
Δ	variation
δ	exponential discount factor, parameter
ℓ	effort
η	decision error
$\hat{\theta}$	estimator
λ, γ	parameters
$\mathbb{E}$	expectation
B	bias
T	test statistic
X	matrix of individual observations, e.g observable characteristics
y	vector of the observations on the outcome variable
$\mathcal{I}$	beliefs in bayesian estimation
$\mathcal{L}$	sampling distribution
$\mathcal{N}$	normal distribution
$\mathcal{S}$	state space
$\mathcal{T}$	treatment

$\mathcal{X}$	inputs
$\mathcal{Y}$	outputs
μ	mean
Ω	variance-covariance matrix
$\omega()$	probability weighting function
B	Binomial distribution
dCor	distance correlation
dCov	distance covariance
F_l, F_u	critical values of the Fisher distribution
Φ	standard normal cumulative distribution
ϕ	standard normal density
π	profit
$\boldsymbol{\varepsilon}$	vector of error terms
ρ	Pearson correlation coefficient
σ, ψ	standard deviations
Θ	parameter space
$\mathbb{V}$	variance
ε_i	individual error terms
a, b, A, B	general purpose parameters (actions, prizes, bids…)
b_L	lower bound of confidence interval
b_U	upper bound of confidence interval
c	threshold in hypothesis testing
$c_e()$	cost of effort
d_0, d_1	decisions in hypothesis testing
DR	decision rule
e	endowment
$F(), f()$	functions
$G()$	cumulative distribution function
$g()$	density
h, i, j, k, s, t	indexes
H_0, H_1, H_a	statistical hypothesis
K	number of samples, treatments, classes…
$L()$	likelihood
$LL()$	log-likelihood
m	number of observable characteristics, median
N	population size
n	number of observations, sample size, number of modeling features
$n_{\mathcal{X}}$	number of inputs
$n_{\mathcal{Y}}$	number of outputs
p, Pr	probability
$p_{(k)}$	rank-ordered p-value
q, Q	price, returns
r	rank
rr	rate of return

S^2	sample variance
SS	sum of squares
T	time, date, period
t_α	critical value of the Student t distribution
$U(), V()$	preference functionals
$u(), v()$	utility functions
w	wage
X, Y	random variables
x, y	realization of random variables
$Y_{(h)}$	ordered value of Y (with order h)
Z	dummy variable
z_α	critical value of the normal distribution
α	Type I error
β	Type II error
θ	parameter(s)
E	event
p_τ	tremble
R	rejection region in hypothesis testing
$W()$	event weighting function
x_{ij}	observation for subject i and variable j
y_i	observation on the outcome variable for subject i

Preface

There is an experimental-economics paradox. Inside the community of researchers carrying out laboratory experiments, these latter are seen as no more and no less than a tool for empirical research. From the outside, however, the method is often perceived as part of a particular sub-field, behavioural economics, which applies insights from both economics and psychology for the better understanding of economic behaviour. Experimental economics is also usually taught this way in most programmes, as part of behavioural-economics classes.

It has, however, long been recognised that experimental and behavioural economics are not the same. Behavioural economics is a research programme with a clear ambition and a well-defined objective: improving economic analysis using realistic psychological assumptions about human behaviour. Experimental economics, on the contrary, is not, per se, a research programme. Rather, it is a research method based on experimental control, applied to the typical topics in economic analysis.

The aim of this textbook is to help close the gap between the perception and reality of experimental methods in economics. We cover experimental economics, i.e. controlled experiments used as a tool to provide empirical evidence that is relevant for economic research. The structure of the textbook thus mimics the way many econometrics textbooks have been written for decades: the coverage focuses on applied statistical methods, the use of which is illustrated with economic results.

There are, however, a number of (good) reasons for this confusion between behavioural and experimental economics, which is at the heart of the experimental–economics paradox. First, behavioural economics emerged partly from the use of experiments – although the contribution of early experiments (such as the Allais paradox and the Chamberlin and Smith market experiments, described in Chapter 1) was to both behavioural economics and mainstream economics (for instance, neoclassical market analysis). Second, the experimental economics method is particularly suited for the study of the phenomena of interest to behavioural economics. In a nutshell, control offers researchers a way of identifying departures from the neoclassical explanation of behaviour. Third, not only behavioural economics but also experimental economics owe a great deal to the accumulated knowledge in experimental psychology: controlled experiments have been used for a long time in this field, and most methodological discussions took place before they even appeared in economics. In addition, the

experimental method is taken as part of the psychology research toolkit across the whole community of researchers.

The scope of this book has been greatly influenced by the place that experimental economics occupies between neoclassical economics, behavioural economics, psychology and statistics. First, our methodological discussion mainly focuses on the use of experiments to understand economic behaviour. We complement this fairly standard view in applied economics by regularly devoting space to insights from, and some discrepancies with, psychology. We also cover a number of standard experimental results that are generally seen as part of behavioural economics.

Second, we mainly focus on laboratory experiments rather than field experiments or randomised controlled trials (see Chapter 3, Section 3.5 for the definition of these). This restriction reflects at least three factors. First, one textbook cannot suffice to embrace the large literature on methods for both laboratory experiments and randomised controlled trials. Second, this restriction also comes from our own limitations in expertise. Last, but not least, laboratory experiments are a convenient step in the study of controlled experiments in economics. Laboratory experiments can be seen as an extreme case of controlled experiments; they allow the accurate identification of behavioural phenomena, but at the cost of a highly artificial environment. Due to this artificiality, laboratory experiments provide answers that are sometimes hard to interpret – and are often challenged by non-experimentalists. Other kinds of experiment offer a way of loosening these limitations by implementing the same empirical method in less artificial contexts. We thus believe that laboratory experiments are a good starting point for anyone who wants to learn about controlled experiments in economics. Many of the discussions in this textbook aim to clarify the most appropriate cases for each type of empirical method; for example, whether observational or experimental data are required and, if it is experimental data, how close to the field the experiment should be.

Structure of the book

This textbook is not the first experimental-economics book by a long way, with respect to both methods and applications. Our predecessors can be split into two groups. First, textbooks/handbooks written for students and academics provide extensive surveys of experimental results. This applies to the textbook of Friedman and Sunder (1994) and the two seminal handbooks edited by Plott and Smith (2008) and Kagel and Roth (1995). In the same spirit, a number of books propose reviews of existing results from laboratory experiments with more specialised perspectives: Camerer (2003) contrasts behaviour in the lab with predictions from game theory, Cartwright (2011) and Chaudhuri (2009) mainly focus on social preferences and behavioural economics, and Angner (2012) provides a detailed overview of laboratory experiments regarding decision problems. These are all required reading for anyone wanting to learn more about experimental results. On the other hand, a few advanced books on the methodology of experiments have recently appeared. These are state-of-the-art collections of papers, written mainly for

academics working in the field. This is the case for Guala (2005), Bardsley et al. (2009) and Fréchette and Schotter (2015).

This textbook is an attempt to build a bridge between these two kinds of reference: it provides a detailed presentation of the methodological aspects of economic experiments for readers (students, academics and professionals) who want to enter the field. To this end the book inverses the usual way of presenting the material, as the experimental results are used to illustrate methodological issues – rather than spreading out the methodological discussions over the presentation of various experimental designs. The content of the book is set out at the end of Chapter 1. We are aware that 'Methodology, like sex, is better demonstrated than discussed, though often better anticipated than experienced' (Leamer, 1983, p. 40). Mimicking the approach in applied economics and econometrics textbooks, the concrete applications of the method that constitute the core material in existing textbooks are here introduced as illustrations of the main material. To this end, the book contains three types of side material describing particular experiments, results or designs: case studies, illustrations and focuses.

- **Case studies** are sections devoted to the detailed presentation of a particular strand of experiments. They seek to illustrate the methodological discussions provided in the corresponding chapter – identified as such in the table of contents.
- **Illustrations** are boxes providing a presentation of one particular experiment or result, to illustrate the point discussed in the text. Illustrations are often provided in sequences, showing how the literature has evolved according to the different dimensions discussed in the text.
- **Focuses** are boxes providing a more detailed and/or formal presentation of a point discussed in the text.

These together provide examples of most of the applications or results that are generally seen as essential in the field – as described in Section 1.4. To help readers bring together all of the information on one particular topic, they appear as specific index headers (see p. 431).

Audience

There are three natural audiences for this book. Its first purpose is as part of a graduate course, describing methods in experimental economics. The organisation of the book closely follows the typical outline of an 8 × 3-hour course. Chapters 1–4 cover the material that would serve as an introductory lecture to laboratory experiments. These chapters describe the main objectives of laboratory experiments and provide examples. Chapters 5 and 8 provide core methodological insights that would best be split in two lectures each. Longer classes could include a discussion of the statistical analysis of experimental data based on Chapter 7 and a discussion of the insights drawn from behavioural economics in Chapter 9, and/or use case studies to devote some lectures to applications that illustrate the main material. In particular, a thorough methodological

course would probably feature some lectures devoted to risk preferences (Section 7.4), time preferences (Section 6.6) and belief-elicitation methods (Section 5.6).

Second, the book more generally seeks to provide future experimental practitioners with a broad picture of the toolkit that they will need. By providing the rationale for the general method and setting out in detail each particular choice of design feature, we hope that readers will be able to construct experiments that fit their research question well. A good understanding of the methodological challenges is also an important requirement for becoming an informed reader: this book may help to interpret the results from laboratory experiments or the writing of referee reports on papers using the experimental method. Third, we hope the community of academics who are new to this literature will find it a useful summary of the current state of the art about what experimental economics can tell us, and under which conditions it provides valuable answers to research questions in economics.

Acknowledgements

The book was written using the course material for PhD/master 2 courses in a number of different places, and in particular at our home institutions. We are more than grateful to the students who attended these classes for their commitment, remarks, scepticism and enthusiasm. We gratefully acknowledge the support from the Institut Universitaire de France.

It is likely that the book would never have reached its final stage without the encouragement, help and remarks from, and discussions with, Jay Shogren. The writing process took such a long time that we will certainly omit many people whose contributions at earlier stages were much appreciated. This also meant that we have worked with many research assistants, whose help very often exceeded what was expected. Our thanks to Lisa Simon and Solene Delecourt for their work on early drafts of some of the chapters; Sophie Cottet for producing the graphs and figures; and Alberto Prati, Guillaume Royer and Shaden Shabayek for their work on some of the boxes. Last, an incredible number of PhD students and colleagues spent a great deal of time reading the first drafts of different parts of the book and provided us with invaluable feedback. We gratefully thank Arthur Attema, Aurélien Baillon, Han Bleichrodt, Aurélie Bonein, Elias Bouacida, Béatrice Boulu-Reshef, Arthur Charpentier, Paolo Crosetto, Laurent Denant-Boémont, Antoine Hémon, Justine Jouxtel, Antoine Malézieux, Elven Priour, Kirsten Rohde, Angelo Secchi, Benoit Tarroux and Adam Zylbersztejn.

While the field of behavioural and experimental economics is sometimes described as over-competitive, it is also one in which researchers from all over the world cooperate on methodological and bibliographic issues, thanks to the ESA discussion group: the discussions there provided us many insights and ideas for which we gratefully thank all contributors. Our gratitude also goes to Sandra Freeland and Andrew Clark for their thorough proofreading of the manuscript, and the editorial team at Cambridge University Press, Phil Good, Neil Ryan and Chris Harrison, for their continuous support and outstanding work.

Part I

What Is It? An Introduction to Experimental Economics

1 The Emergence of Experiments in Economics

> There is a property common to almost all the moral sciences, and by which they are distinguished from many of the physical; that is, that it is seldom in our power to make experiments in them.
>
> Mill (1836), cited in Guala (2005, p. 2).

This statement by John Stuart Mill, or similar remarks, introduces virtually all texts on the methodology of experiments in economics. At the time, and for a long time after that, controlled experiments in the social sciences, and especially in economics, were considered impossible to conduct; it appeared that experiments were reserved to the natural sciences, and that the testing of social and human behaviour in the framework of a controlled experiment would prove completely unworkable. Nowadays, experiments are a widely accepted means of generating knowledge in economics. Among many examples, it is shown by the fact that experimental or behavioural economics is part of the graduate programme of most universities, there are many books, handbooks and textbooks focusing on the field, and even a well-recognised academic journal ('*Experimental Economics*') is specialised on research using this method.

Before moving on to a detailed discussion of why and how laboratory experiments are performed in economics, we will explore this intriguing trend. What happened between the time experimental economics first came into existence and when it finally became an established member of the community? We will start by highlighting the progress of experimental methods in economics, from an area that was thought impracticable, meaningless or uninteresting, to an accepted and widely used process in economic research. In describing the reasons why there was such a sudden change of interest in and attitude towards experiments, we will examine some of the very first examples of experiments in economics. These examples are interesting not only from a historical point of view, but also because they underscore the main reasons for the change and how experimental economics has grown since – both in terms of the research questions that are addressed and in the type of answers it provides. These will be followed by three more recent examples which illustrate what the research programme has become today – a unified and also very diverse area of study.

The most obvious and powerful unifying factor of all works using laboratory experiments is, in fact, the methodology applied: a controlled environment allowing use of the observed behaviour of human beings to produce knowledge about economics. As the last section will show, a thorough study and presentation of this methodology requires

a wide-ranging knowledge of economic theory as a whole, and its relation to different application fields, analytical tools and approaches. It will soon become clear that no single textbook can possibly cover all these aspects: this chapter will offer a road map of everything this book is unable to cover, or can only cover in part. Perhaps more importantly, this chapter will try to convince you that in order to fully understand the rationale, contribution and practical lessons of the results generated by experiments in economics, the first step is to be aware of the choices of methodology and the reasoning behind them: this is what this book is all about.

1.1 The End of a Long-Standing Regretful Impossibility

Even if experiments in economics were considered impossible for a long time, they were nonetheless the object of considerable wishful thinking. If experiments could be implemented, they could be designed and put in place in order to provide empirical evidence and serve as a basis to enhance theory. This is implicitly acknowledged in a celebrated remark made by Friedman, 'We can seldom test particular predictions in the social sciences by experiments explicitly designed to eliminate what are judged to be the most important disturbing influences' (Friedman, 1953, p. 10). Experiments in the social science are seen as a very attractive, though impossible, way of testing theories. If feasible, experiments would allow researchers to neutralise all forces driving behaviour that are outside the scope of the theory. In that case, experiments would help elicit the empirical content of theory, and therefore identify the main driving forces of behaviour. This opinion was shared by many eminent economists long after 1953. In their groundbreaking principles textbook, Samuelson and Nordhaus noted that 'economists cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors' (Samuelson and Nordhaus, 1985, p. 8). All of the remarks cited above show quite clearly how recent the appearance of experimental economics as a *bona fide* field of study is and also underline how desirable experiments are for research. Fortunately, the long-standing and powerful belief in the impossibility of experiments in the social sciences, however regretful, is now a thing of the past.

As a matter of fact, in a later edition of their textbook (which appeared less than ten years later) Samuelson and Nordhaus had already adopted a new and different mindset: 'Experimental economics is an exciting new development' (Samuelson and Nordhaus, 1992, p. 5). Between these two editions, economists had managed to set up experiments similar to the ones conducted in the natural sciences. But, even more importantly, the results generated by these experiments began to be considered by an increasing number of specialists to be sound empirical evidence.

From then on, the pace and scope of the changes taking place increased so rapidly that today the situation stands in sharp contrast with the earlier views expressed above. This phenomenon is illustrated, for instance, by the rise in the rate of academic publications related to experimental economics over the years. Figure 1.1 shows the results of a survey carried out by Noussair (2011) concerning the percentage of articles including experiments that have appeared in major academic economic journals. The survey

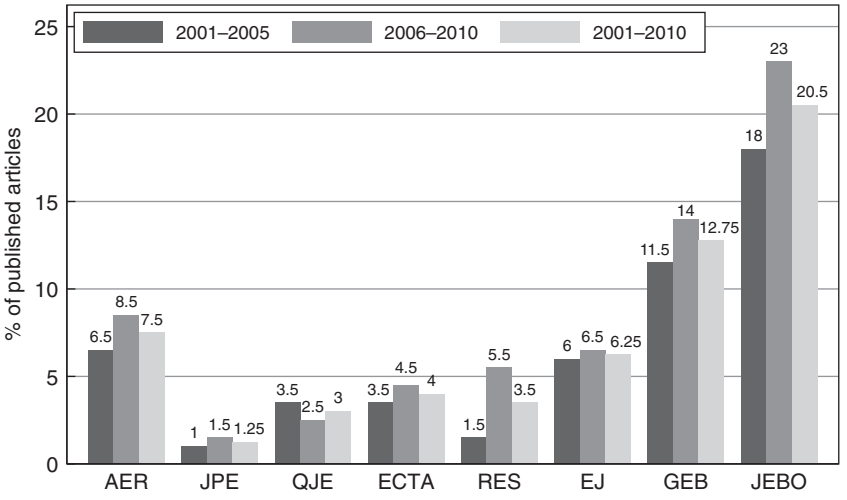


Figure 1.1 Trends in academic publishing in experimental economics

Note. Percentages of experimental articles from those appearing in the journals: *American Economic Review* (*AER*), *Journal of Political Economy* (*JPE*), *Quarterly Journal of Economics* (*QJE*), *Econometrica* (*Ecta*), *Review of Economic Studies* (*RES*), *Economic Journal* (*EJ*), *Games and Economic Behaviour* (*GEB*), *Journal of Economics, Behavior and Organization* (*JEBO*).
Source: Noussair (2011, p. 8).

covers the top five journals (*AER*, *JPE*, *QJE*, *ECTA*, *RES*) which experts acknowledge as the leading supports in the field; three other journals were added to the list: *EJ*, *GEB* and *JEBO*. These are more specialised and/or lower-ranked journals, but which are, nonetheless, highly influential and open to experimental works. The chart shows the change in the rates from 2001–2005 to 2006–2010. The first ten years of the new millennium saw a slight increase in the percentage of articles in the sample. More importantly, the share of experimental papers is very significant in most of these leading journals: from 2% to 7% in the top five journals, and from 5% to 20% in the more specialised ones. This a clear indication of the growing acceptance and recognition of this type of work by the academic community.

The four experimental economists who have been awarded the Nobel Prize in Economics in the first decades of the new millennium, who we will come across a number of times in this book, are another example of this recognition. In 2002, Vernon L. Smith and Daniel F. Kahneman were the joint recipients of the Nobel Prize in Economics. Smith was thus acknowledged as one of the founders of experimental economics and as someone who contributed to establishing it as a conclusive method. The main justification for the award was the introduction of the methodology per se (they received the prize 'for having established laboratory experiments as a tool in empirical economic analysis, especially in the study of alternative market mechanisms'). In terms of contributions, the field is seen as interdisciplinary in nature, with Kahneman receiving the prize 'for having integrated insights from psychological research into economic science, especially concerning human judgement and decision-making under uncertainty'. Ten years later, another renowned experimentalist, Alvin Roth, was also granted the

Nobel Prize. But this time, the co-winner was Lloyd Shapley, a pure theorist. Together they were recognised 'for the theory of stable allocations and the *practice* of market design'. It goes without saying that the Smith and Kahneman contributions are of major importance to the discipline, and that these three Nobel Prizes in themselves are convincing proof that experiments have been widely accepted as part of the field. But there is an interesting change in nature between the two prizes: while the first Nobel Prize was awarded for the methodological advance itself, the acknowledgement of Roth's contribution was based on actual laboratory results using the toolbox of experimental economics and applied to research issues that are at the core of economic theory. This is further evidence of the wide acceptance of experimental economics by the academic community. Last, Richard Thaler was awarded in 2017 for having incorporated 'psychologically realistic assumptions into analyses of economic decision-making'. Richard Thaler showed how experimental methods are particularly meaningful for uncovering deep psychological phenomena such as mental processes, self-control behaviour and social preferences. The award also underlines his contribution to public policies based on nudges (see Chapter 9). This is further evidence of the wide acceptance of experimental economics by the academic community, with results from the laboratory now being seen as useful in order to better design choice architectures.

In contrast with the quotes that opened this section, in which experiments were regarded with substantial scepticism, there is now substantial evidence that experimental economics has become a well-established and widely accepted empirical method. One may wonder how an entire new field has managed to surface in such a short period of time. As a first step towards a better understanding of how this change came about, we will show in the next section that this, in fact, was not the case at all: experiments in economics have existed for a long time, producing results that are much in line with the works that appear nowadays in leading publications. It appears that the reason for the lack of experiments in economics comes not so much from their practical impossibility, but rather from the main focus of academic research at the time. Since then, a change in focus occurred towards questions that are closer and closer to the kind of issue that experiments are well suited to investigate.

1.2 Why Such a Change: Two Early Examples

The two examples below are among the best known of the early experiments. They illustrate the state of infancy of experimental economics at the time, although they are now regarded as important and insightful contributions to economic knowledge.

1.2.1 How Do Competitive Markets Work?

In 1948, Harvard Professor Edward Chamberlin organised a game with his students. The aim was to replicate the functioning of a market in perfect competition with rational agents as closely as possible. Students were randomly assigned a card, which made each student either a seller or a buyer. In addition, the card displayed a price for a hypothetical

good to be sold or bought. For students playing as sellers, this price referred to the minimum price at which they were ready to sell. For the buyers, this price indicated the maximum price they were willing to pay to obtain the (hypothetical) good. Afterwards, the students walked freely in the classroom and bargained with their colleagues to either buy or sell the good. Once a deal had been made, the students came to Chamberlin's desk to report the price at which the good had been sold.

In this framework, economic theory predicts outcomes according to the two curves depicted in Figure 1.2, where the supply and demand curves were drawn based on the prices distributed to students – i.e. how many students were willing to buy or sell at each possible price that appears on their card: a 'induced values' design. The game is a textbook example of a market: the demand curve is decreasing in price, whereas the supply curve is increasing. The market equilibrium determines the actual price that should arise from strategic interactions, as well as the resulting quantities exchanged on the market; the unique stable price is the one that clears the market, in such a way that demand meets supply. This point is an equilibrium not only because the two sides happen to be equal, but more importantly because it is the only state of the market in which everyone agrees to stay – there is no possibility of doing better at the individual level by moving out of this situation. For any other price, there is either excess supply or excess demand, in which case either suppliers (sellers) or consumers (buyers) can be in a better situation by moving to another price level. There are thus strong reasons to believe that the equilibrium should result from real interactions in this particular environment.

Surprisingly enough, Chamberlin obtained the results reported in Figure 1.3 based on the actual behaviour of his students. The dashed line depicts the average price at which students traded their goods during the experiment: it is far below the straight line, or the competitive equilibrium price. There was also a huge variation in the actual prices,

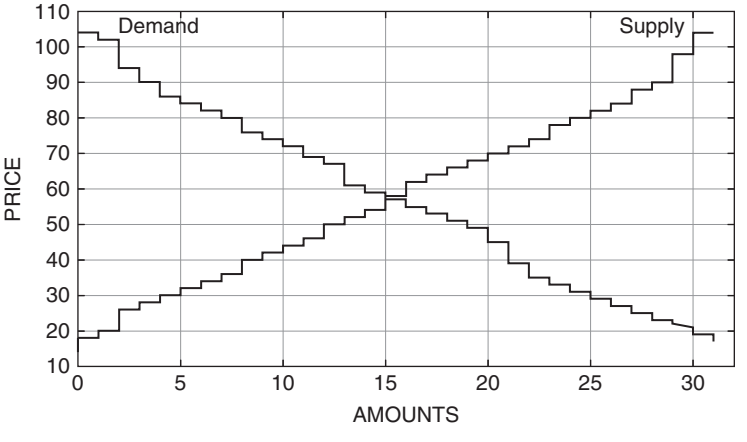


Figure 1.2 Market equilibrium in the Chamberlin (1948) experiment

Note. The figure shows the theoretical equilibrium of the market implemented in the laboratory – at the intersection of the (increasing) supply function and the (decreasing) demand function.
Source: Chamberlin (1948, p. 97, Figure 1).

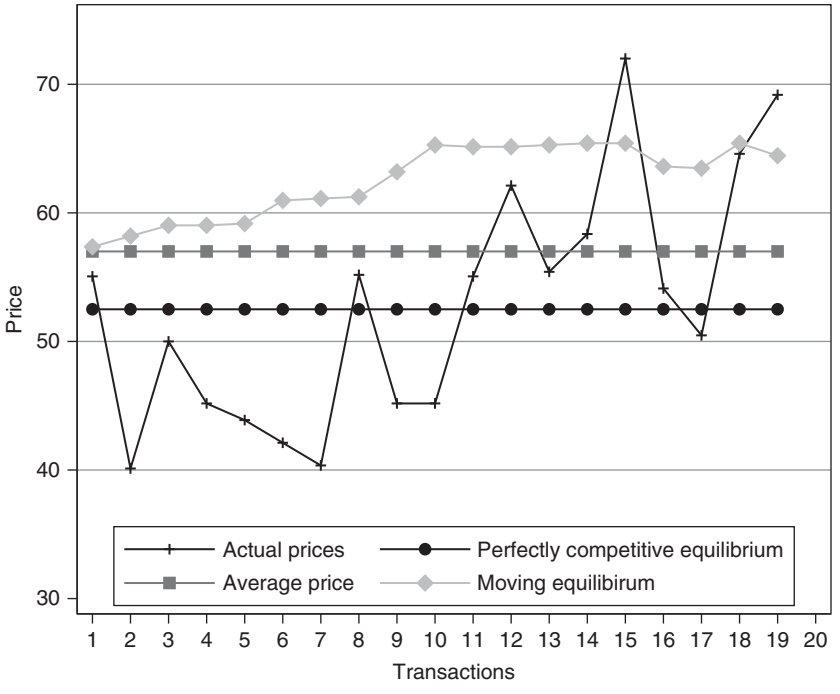


Figure 1.3 Observed behaviour in the Chamberlin (1948) experiment

Note. For each transaction in abscissa, the figure shows the actual price observed in the experiment as well as a recall of the theoretical equilibrium described in Figure 1.2.
Source: Chamberlin (1948, p. 101, Figure 3).

which are represented by the curving line. In addition, the equilibrium volume of trade is higher than what the theory would have predicted. Actual behaviour in this environment thus strongly departs from what economic theory expects, leading Chamberlin to conclude, 'Perhaps it is the assumption of a perfect market which is "strange" in the first place' (and interpret this as a support for his monopolistic competition model). This result is not, however, the end of the experimental story of markets.

Vernon Smith (who, as mentioned above, was subsequently awarded a Nobel Prize) was one of Chamberlin's students and participated in his classroom experiment. Around fifteen years later, in 1962, he decided to replicate Chamberlin's experiment, but with various changes in the environment – aimed at replicating what Smith thought were important actual driving forces of a competitive market. As in Chamberlin's experiment, each student received a card, making him either a buyer or a seller. This card also gave the student a reservation price: the price above which a buyer would not buy, and below which a seller would not sell. The changes implemented as compared to the seminal experiment are as follows. First of all, instead of having bilateral bargaining (or, at most, discussions in small groups) between students, the announcements of offers and demands become public, meaning that buyers and sellers could call out their offers in the room so that everybody could hear. This is aimed to make the information on prices public, so as to mimic what is achieved by an auctioneer receiving and distributing all

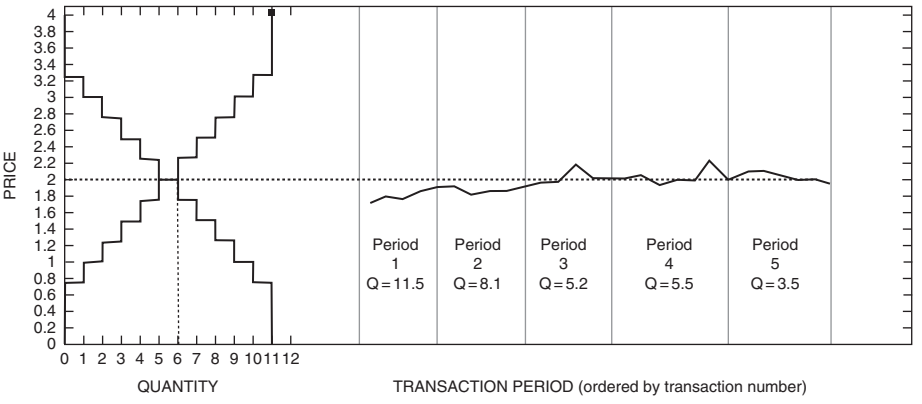


Figure 1.4 Predicted and observed behaviour in the Smith (1962) replication

Note. The left-hand side shows the theoretical market equilibrium – at the intersection of the (increasing) supply function and the (decreasing) demand function. The right-hand side shows the price and number of transactions in each market period.

Source: Smith (1962, p. 113, Figure 1).

offers. Second, the market experiment was repeated over several periods, and allowed the students to better understand the functioning of the market, hence getting closer to market behaviour of professional market traders.

Figure 1.4 reports the observed behaviour and theoretical predictions of the Smith experiment. The theoretical market plotted on the left-hand side shares the same features as the one implemented by Chamberlin. The curve on the right-hand side shows the prices at which market clears for five market periods. The contrast with the previous results is drastic: the observed prices smoothly converge towards the equilibrium price, and the number of transactions (reported on the bottom part of the graph) converges to the equilibrium quantity equal to 6.

Beyond the seminal insights about how the market works, these series of experiments help to describe the methodological issues behind experimental results. Both experiments aim to replicate competitive markets, but with different implementation choices. The best environment to describe markets is a matter of judgement, and the theoretical conclusion drawn will be entirely different whether one or the other experiment is believed to best capture the important features of the economic phenomenon. At the same time, the implementation differences between the two experiments also inform about the key features that explain behaviour in a market situation: the extent of information buyers and sellers receive, for instance, seems to be a critical driving force. Beyond rejection/support of the prediction, the experiment thus informs theory by highlighting the salient dimension to be taken into account. Lastly, as the Smith experiment clearly shows, it is not always the case that the theory is necessarily wrong or that experiments are designed expressly to reject the behavioural assumptions behind the theoretical results (as is sometimes taught, mainly by some academics who view experiment results with scepticism): in this case, experiments serve more to identify the circumstances under which these assumptions are actually accurate.

Table 1.1 The choice sequence of the Allais paradox

	Option A	Option B
A or B?	100% chance of winning 1 million	10% chance of winning 5 million 89% chance of winning 1 million 1% chance of winning nothing
	Option C	Option D
C or D?	11% chance of winning 1 million 89% chance of winning nothing	10% chance of winning 5 million 90% chance of winning nothing

Note. Each respondent was asked to make both choices in turn.
Source: Allais (1953, implemented in 1952).

1.2.2 Choice Consistency in Risky Decisions

The second example focuses on individual decision-making, rather than on strategic situations. During the annual conference of the American Economic Society held in New York City in 1953, Maurice Allais presented the economics professors attending the conference – especially those specialised in game theory and decision theory – with two binary choices. Respondents were shown Table 1.1 and asked to choose either A or B, and then either C or D.

Based on the axiomatic framework of decision theory, the first choice and the second choices are strongly related – although the choice between the two options per se is a matter of preferences that nobody can predict. To understand the link between the two decisions, let us first put aside the 89% probability of winning one million – in situations A and B – or nothing – in situations C and D. Apart from this 89% probability, both situations A and C have the same probability (11%) of winning one million. Similarly, situations B and D offer the same expected outcome: nothing with a probability equal to 1%, and five million with a probability of 10%. As a result, still disregarding this 89% probability, an individual who prefers A over B (B over A) should also prefer C over D (D over C). You can note that the outcome that results from the 89% probability is exactly the same for A and B on the one hand, and C and D on the other. Consequently, it only comes down to the addition of an identical outcome for each pair of situations: one million for A and B, nothing for C and D. It sounds reasonable to assume that this should not affect the preference ordering of consistent decision-makers.[1] Because of this very clever feature in the way situations are built, elicited choices provide a test of consistency: depending on individuals' unknown preferences, either A and C, or B and D, should be picked together; no other combination can be rationalised with classical decision theory. Using these choice situations, Allais was successful at tricking the economists at the conference. As he expected, 45% of the leading theorists (including Savage, one of the leading researchers in the field) to whom Allais submitted the choice

[1] This property of preferences is named the "independence axiom" in decision theory, which implies that if there are two different gambles and one is preferred to the other, then mixing them with another identical gamble should not alter the order of the preferences. This axiom is the one violated by the results of this experiment, which is now known as the common consequence or Allais paradox.

opted for A against B, but D against C. Almost half of the respondents, who were all well versed in economic and decision theory, and some specialised in decision theory, failed to pass the consistency test associated with the two successive choices. A key feature of this experiment is that it is designed in such a way that there is a unique relationship between one, clearly identified, theoretical assumption driving the predictions and the choices available. Therefore, observed behaviour challenges not only theory, but, more importantly, the specific feature of theory that fails to describe behaviour. Beyond simple rejection (which is unambiguous given the magnitude of the result and the sample pool from which it was obtained) it provides a guide to the particular assumptions that have to be reworked so that they correspond to the real driving forces behind behaviour. In the Allais paradox, two features of the available options are of particular interest. On the one hand, certainty generates a strong attractiveness for option A. On the other, the change in probabilities appears to be quite small between options C and D. These two features of behaviour under uncertainty are central in theories that rationalise behaviour in the Allais paradox (Quiggin, 1982; Kahneman and Tversky, 1979).

1.2.3 Why Was There Such a Fast and Sudden Change?

These early experiments marked the beginning of a new field, which has made rapid gains in terms of both acceptance and popularity over the last decade. But many years went by between the time of those first experiments and the time when the economic community truly started paying attention to them. Until recently, experimental economics was thought of as unworkable or of no meaningful importance. What was it, then, that suddenly made the experimental method so widely accepted?

As shown by the two previous examples, this was not a matter of feasibility. Both experiments were published in very good journals and existed when some of the quotes opening this chapter were written. Experimentation was thus already a possibility. In fact, it has always been quite straightforward to test results from decision theory or game theory in an experimental setting. It simply amounts to having people make choices within a simple set of rules describing the decision-making environment. The most drastic change was in fact the change in the kind of questions, which in the 1970s and 1980s economics began to focus on, with a growing importance put on these two theoretical tools.[2] In the middle of the twentieth century, economics was set in the context of a beautiful model of how the entire economy worked and how all the agents in the economy, as a group, made decisions in the present and for the future. This environment was so complex and all-encompassing that the empirical relevance of behavioural assumptions was obviously not a primary concern. But as economics moved away from this representation, more and more attention began to be given to the forces behind individual and strategic decision-making. Microeconomics became one core focus of economic analysis, making an intensive use of game and decision theory. What were considered revolutionary issues at the time have now become orthodox, and the rise of experimental economics was concurrent with the fall of general equilibrium theory. The reinforcing of

[2] See Fontaine and Leonard (2005), in particular Chapter 3, for an insightful review of these trends.

the economic representation of human behaviour, along with clear-cut definitions of the environment, has now made the long-held dream of testing economics in the laboratory an achievable goal.

The role of experiments in the history of economics helps better understand what experimental economics is all about. First, experiments and economic theory go hand in hand: experiments are about assessing the empirical relevance of the behavioural content of economic models. They are not in contradiction with economic theory, but rather serve as a complement to it. Economic theory provides a deep and subtle understanding of how the economy works when decisions are taken by the *homo œconomicus*. Experiments rather involve a *Homer œconomicus*: the ordinary Joe, endowed with an average level of cognitive and social skills – rather than unlimited computational abilities – under the influence of psychological and environmental factors – rather than driven by a well-defined preference functional.[3] They thus allow us to measure whether *homo œconomicus* and *Homer œconomicus* lead to similar or substantially different outcomes in a given economic situation. Second, as the two examples cited above show, these two kinds of people, the *homo œconomicus* and 'real' human beings, are not strangers to one another: they sometimes behave differently, calling for a different theory (rather than different people), but there are also many important situations in which the two behave as if they were one and the same. Why and when they do is one of the key questions that remains to be answered.

1.3 The Research Programme: Three Examples

We conclude this quick overview of the recent history of experiments in economics with three examples drawn from a more recent literature. Although chosen at random (with bias) among many other similar studies, these examples clearly illustrate the current state of the art in the experimental field, and the way it helps elucidate what human beings and economic theory – *Homer œconomicus* and *Homo œconomicus* – have in common, and how they differ. To a large extent, the current answer is similar to the main lesson we learned from the early examples described above.

We first present the prisoners' dilemma (PD), a well-known example of the discrepancy between game-theoretic results in a simple environment, and the behavioural patterns actually observed. This example also shows that, while game theory alone has trouble explaining behaviour – typically, without reference to more general factors related to economic agents' social environment – it is in fact quite effective in predicting changes in behaviour. The second example shows that experimental economics can help significantly in this aspect as it can easily address difficult questions about the basics of economic rationality. The centipede game is a typical example of a simple experiment that calls into question some common principles of rationality. Lastly, we proceed to a more complicated game, a zero-sum game with incomplete information, in which one would expect the gap between economic theory and observed behaviour to be larger

[3] The terminology is due to Hall (2005); see e.g. Beggs (2013); Hall (2014) for a full statement of the parallel between economics and the Simpsons.

than in any other context. The example shows that this foregone conclusion is definitively not applicable in this case. These three examples serve as a tour of the type of research question addressed thanks to experiments in economics and of the variety of answers it offers.

1.3.1 Nash Equilibrium and Pareto Efficiency

The first example is that of a *non-cooperative game*: a game in which outcomes are determined by the decentralised and independent actions of players. Figure 1.5 presents the payoffs each of the two players gets according to the actions they choose. It is a simultaneous-move game, as each player decides without knowing what the other one is doing. This type of representation of a game will be used often in this book. For readers who might not be fully familiar with it, we will take the opportunity here to describe it step by step.

According to the normal form representation in Figure 1.5, each player can choose between two actions. Player 1 is the row decision-maker, and Player 2 is the column decision-maker. Player 1 chooses either Top or Bottom, Player 2 either Left or Right. Together, both players' actions determine the outcome of the game: the state of the world resulting from all the players' actions. The numbers in the matrix show the payoffs linked to each of the four possible outcomes for each player. In each cell, the number on the left is the payoff Player 1 gets in this particular outcome, and the one on the right the payoff Player 2 gets. For example, if Player 1 plays Top and Player 2 chooses Right, then Player 1 loses 10 and Player 2 earns 10.

A quick inspection of Figure 1.5 shows that one outcome seems intuitively preferable: if the players choose Top of Left, they reach an outcome that maximises what they collectively get. It is a Pareto-dominant outcome: that particular outcome makes it impossible for one agent to improve his lot by unilaterally modifying his action without making the other player worse off. However, this outcome is not sustainable when the actions are decentralised and non-coordinated. This is so because, given the Pareto-dominant situation, both agents have an incentive to deviate: given the action of Player 2, Player 1 can earn more by playing Bottom than Top against Left, and similarly Player 2 can earn more by playing Right against Top rather than Left. Because of these individual incentives to move away from the Pareto-dominant outcome, the equilibrium coincides with the worst outcome of the game: that which occurs when Player 1 chooses Bottom and Player 2 chooses Right. This is a Nash equilibrium because there is no longer any

	Left	Right
Top	5; 5	–10; 10
Bottom	10; –10	–5; –5

Figure 1.5 Table of payoffs in a non-cooperative game

individual incentive to deviate – none of the players can be better off by moving away from the equilibrium strategy when the others are playing it.

The Nash equilibrium of the game, Bottom–Right, is the outcome that is predicted to occur from uncoordinated simultaneous decisions. Because it does not coincide with the Pareto-dominant situation, this game is a textbook example of the failure to reach an efficient outcome via non-cooperative decisions. It is often called the prisoners' dilemma game, in which case the moves are 'to denounce' or 'not to denounce' for two prisoners who are separately offered leniency if they provide information about the crime they committed together. This strategic framework can be applied to a great many economic situations. Collusion between firms on markets is a typical example of the dilemma of cooperation and defection (which will be studied in length in Chapter 4, Section 4.4.2). Collusion occurs when firms agree to set the market price to a level higher than its competitive value. All firms prefer the collusive outcome, as profits are higher. But each firm has a strong temptation to slightly decrease its price so as to make even higher profits, at the expense of others. This incentive to deviate from the collusion agreement is a natural force against the ability to sustain a non-competitive equilibrium. Another example of non-cooperation when cooperation would be optimal is the Kyoto Protocol, an international agreement which aims to commit countries to reducing their greenhouse gases. A Pareto-optimal outcome would be that all countries sign the agreement. Nonetheless, countries have an incentive to let the other countries sign and to free-ride, thus benefiting from the reduction in greenhouse gases without having to pay the price of the treaty.

Hundreds, if not thousands, of experiments have been run to assess the empirical relevance of this analysis. As an example, Figure 1.6 presents the results of one of the earliest experiments of this type, conducted by Cooper et al. (1996). The x-axis represents each of the ten different periods of the game, while the y-axis depicts the frequency of the cooperative play (i.e. when the collectively optimal, but not individually rational, actions are chosen) when the action leading to the efficient outcome is chosen. The upper curve represents the frequency of cooperative play in the case of a prisoners' dilemma game with repeated interactions, where the same two players play together ten times. The lower curve represents the outcome with different partners for ten periods, each game being a one-shot game. Both curves show a departure from theoretical predictions. Theory predicts a 0% rate of cooperation in the game. It is far from the observed patterns not only in the repeated games – which do not, in the strict sense, implement the model – but also in the one-shot games. For example, in the first period, about 60% of the subjects decided to cooperate in the case of finitely repeated games, but around 35% of the people did so in one-shot games. At the same time, it is not true that these results fit with a view of human behaviour only driven by the well-being of everybody and disregarding self-interest. Free-riding behaviour, based on the temptation to increase one's payoffs at the expense of the other players, accurately describes the results in 70–50% of observed outcomes. Because these two kinds of behaviour (cooperation and deviation) are widespread, both should be accounted for by any accurate theoretical representation. As a result, neither the Nash equilibrium, nor alternative motives leading to full cooperation, are enough to account alone for the observed behaviour in the prisoners' dilemma game.

Figure 1.6 Empirical behaviour in prisoners' dilemma games

Note. The figure reports the share of participants who decide to cooperate in each of the ten periods of the game.

Source: Cooper et al. (1996, p. 199, Figure 1).

1.3.2 A Simple Two-Player Sequential Game

The previous example focused on a simultaneous-move game, in which the players decide without knowing what the others will be doing. Another branch of game theory studies behaviour in sequential-move games, in which the players decide one after the other. The big change in terms of strategic interaction is that each player now observes what the other did before choosing an action. Figure 1.7 provides a well-known example of such a game, introduced by Rosenthal (1981) as the *centipede* game. The structure of the game is quite simple. Two players alternately get a chance to take the larger portion of a continually increasing pile of money – the number on each node indicates which of the two players has to decide, with the two payoffs being those experienced by each of the two players respectively if the game ends at this point. For Player 1 payoffs are given in the first row, for Player 2 payoffs are given on the second row. The amount keeps increasing as long as the players continue to play (denoted P in Figure 1.7 and Figure 1.8). But as soon as one of the two players decides to take (denoted T in Figure 1.7 and Figure 1.8), they get a larger portion of the pile while the other gets the smaller part. The trade-off is not easy to resolve from an intuitive point of view: conditional on the game continuing, it is always better to go as far as possible along the tree (the original form had 100 nodes, hence the name centipede), but at the same time each player wants to be the one who stops the game. The question that remains open, then, is when the players will stop and at which stage.

The way game theory resolves this trade-off is, in a sense, even less intuitive than this simple explanation suggests. The key point to note is that the number of steps in the

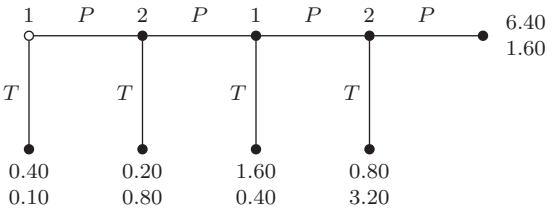


Figure 1.7 A simple four-moves sequential game

Note. Each of the two players (1 or 2) decides in turn at each node to either Pass or Take. For each state, the payoffs of Player 1 appear on the first row, the payoffs of Player 2 on the second row.
Source: McKelvey and Palfrey (1992, p. 806, Figure 1).

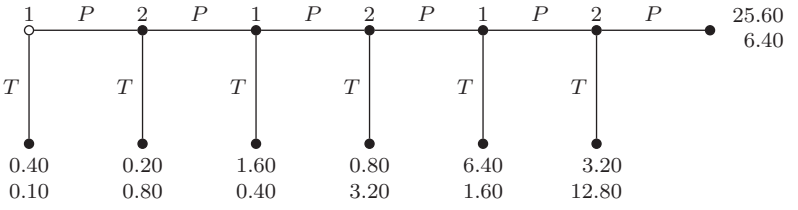


Figure 1.8 A six-moves centipede game

Note. Each of the two players (1 or 2) decides in turn at each node to either Pass or Take. For each state, the payoffs of Player 1 appear on the first row, the payoffs of Player 2 on the second row.
Source: McKelvey and Palfrey (1992, p. 806, Figure 2).

game (four in the example) is known for sure from the beginning. The usual approach to this type of situation is to predict that the players will play in such a way that actions in each sub-game (i.e. the sub-tree that extends from any node to the end) is a Nash equilibrium. Because of this property, the equilibrium can be elicited through backward induction. Starting from the terminal node of the game, the equilibrium behaviour is relatively straightforward: the last player will decide to take, because it earns more than to pass and there is no point in waiting. For Player 1, in the node just before, it means that the decision actually faced is between taking now or having Player 2 take at the last stage. But then the best thing to do is to take at this node so as to avoid letting the other player take at the following one. And this reasoning applies to all the steps leading backward taken one after the other. The result of this reasoning would be the *sub-game perfect equilibrium*, where the first mover takes at the very first node and the game stops. What is startling in this result is that the outcome is not predicted to depend on either the rate at which the pie grows from one step to the other, or on the number of steps – as long as the number is known right from the beginning.[4]

This striking prediction was first tested against actual behaviour by McKelvey and Palfrey (1992). When the participants in their experiment were asked to play the four-move game described in Figure 1.7, only 7% of them actually played according to the sub-game perfect equilibrium, stopping at the very first node. The top part of Table 1.2

[4] See Reny (1993); Aumann (1995); Ben-Porath (1997); Aumann (1998) for theoretical attempts to weaken this paradoxical result, and Chapter 4, Section 4.4.2, for a more detailed discussion of finite- and infinite-horizon games.

Table 1.2 Observed continuation decisions in centipede games

		Session	N	f_1	f_2	f_3	f_4	f_5	f_6	f_7
	1	(PCC)	100	.06	.26	.44	.20	.04		
Four	2	(PCC)	81	.10	.38	.40	.11	.01		
Move	3	(CIT)	100	.06	.43	.28	.14	.09		
	Total	1-3	281	.07	.36	.37	.15	.05		
High Payoff	4	(High-CIT)	100	.15	.37	.32	.11	.05		
	5	(CIT)	100	.02	.09	.39	.28	.20	.01	.01
Six	6	(PCC)	81	.00	.02	.04	.46	.35	.11	.02
Move	7	(PCC)	100	.00	.07	.14	.43	.23	.12	.01
	Total	5-7	281	.01	.06	.20	.38	.25	.08	.01

Note. Actual behaviour in the four-move (upper part) and six-move (bottom part) centipede game. N denotes the number of subjects, each column f_t provides the share of subjects who decide to take at the t^{th} node.
Source: McKelvey and Palfrey (1992, p. 808, Table IIA).

shows the full distribution of the share of subjects who stopped at each node of the game (denoted f_t for the t^{th} decision stage). While it is true that the subjects were not anyway near to playing the sub-game perfect equilibrium, at the same time few of them reached the last stage of the game – less than 5% did. From this evidence, the question remains open: what is it that makes subjects decide to stop or go on?

To help answer this question, McKelvey and Palfrey (1992) consider a second experiment that implements the six-move centipede game displayed in Figure 1.8. The results are displayed in the bottom part of Table 1.7. From 7% in the four-moves game, the share of subjects who play the sub-game perfect equilibrium is now almost 0 – only two subjects out of 281 do so. But again, the distribution of subjects according to the node at which they decide to stop is not concentrated at the end. Half the subjects rather decide instead to stop at node 5 or 6, two steps before the last stage.

From these two examples, it appears that sub-game perfectness clearly fails to predict behaviour in the extremes. But at the same time, this theory accurately mirrors the trade-off people face in this type of situation: as subjects reach a node closer and closer to the end, it becomes more and more difficult for them to maintain a decision to pass, and more and more likely that the decision they will take as the game proceeds is to stop a few rounds (two to four) before the end.

1.3.3 The Use of Private Information

The first two examples were simple games whose results challenge theory in one way or another. As a third example, we will move on to another quite different environment in which both the rules and the strategies are far more complicated. Its full name is a zero-sum repeated game with incomplete information – each part is explained in turn below. Figure 1.9 shows the stage games of two different versions of the game. We first focus on the non-revealing (NR) version of the game – the difference with the fully revealing (FR) version will be described later.

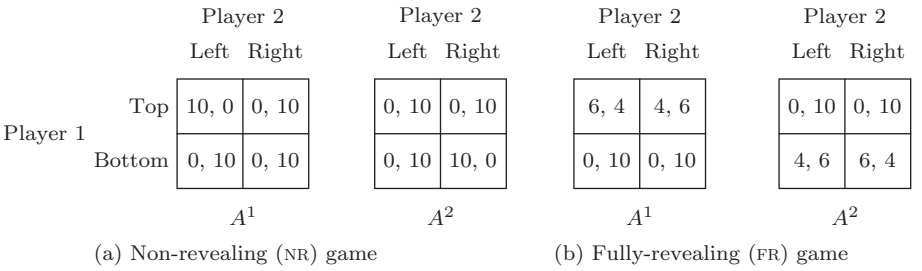


Figure 1.9 Payoff matrices of two zero-sum games

It is a zero-sum game because the payoffs are such that everything that is won by one player is lost by the other – as opposed, e.g., to the prisoners' dilemma game – so that concerns about the situation of other players have no influence over the results. All stage games shown in Figure 1.9 involve two players and two actions. Player 1 chooses either Top or Bottom, and Player 2 chooses either Left or Right – both players decide simultaneously. The numbers in the matrix represent the payoffs of both players after they have chosen their move. In matrix A^1 of Figure 1.9.a, for instance, '10, 0' indicates that Player 1 gets 10 and Player 2 receives 0 if Top/Bottom is played.[5] Inspecting the payoff tables, the game is straightforward to play for both players, because they both have a dominant strategy, i.e. an action which is preferable whatever the action chosen by the other player. For instance, in this matrix A^1, Player 1 does better by playing Top rather than Bottom, and Player 2 by playing Right rather than Left – whatever the decision of the other player. Similarly, in matrix A^2, choosing Bottom is a dominant action for Player 1 and Left is a dominant action for Player 2.

The information structure of the game makes it more interesting than a complete-information zero-sum game. In fact, a random draw (with equal probability) decides on the 'state of the world' before any decision is made. This state of the world is the payoff matrix, either A^1 or A^2, that players are facing. There is incomplete information (on one side) because players are asymmetrically informed about the result of this draw: only Player 1 is given this information. First consider the situation in which the stage game is played only once. Player 1 is privately informed of the consequence of each action and can thus pick up the dominant action of the matrix that has been drawn. Player 2, by contrast, needs to decide without being aware of the state of the world, and will thus randomise between Left and Right. But this information structure in fact becomes interesting when the game is repeated – players face the stage game together several times, and the state of the world is drawn once for all at the beginning. In this context, Player 2 can infer some information about the state of the world from the observed decisions of Player 1.

To see it more clearly, suppose you are Player 1 facing the stage games of Figure 1.9.a and knowing which state of the world you, and the other player, are in. You have to

[5] The sum of players' payoff is positive rather than equal to 0, but since the sum is constant across decisions, it is conceptually equivalent to a zero-sum game.

Table 1.3 Theoretical predictions in the non-revealing and fully revealing games

	Value of the game, v_t if t is						Optimal use of information
	1	2	3	4	5	∞	
FR	5.00	4.50	4.33	4.25	4.20	4	Fully revealing
NR	5.00	3.75	3.33	3.21	3.07	2.50	Non-revealing

Note. Theoretical predictions on behaviour in the NR and FR games.
Source: Jacquemet and Koessler (2013, p. 110, Table 1).

choose between Top and Bottom and you know the game will be repeated. You also know that Player 2 would like to play Right in A_1 and Left in A_2. First imagine that you decide to use your dominant action: you play Top if A_1 is drawn and Bottom if it is A_2. If this is an equilibrium strategy, then Player 2 knows this is how you react to the draw: observing Top delivers perfect information to Player 2 that A_1 has been drawn. At the next stage, Player 2 will thus play Right. But the combination Top, Right is clearly not in your interest, since you get 0: by revealing your information you no longer benefit from it. The other options for you are either not to use your information at all (deciding with equal probability between the two decisions as if you did not receive the information about the draw) or to use it only slightly, by playing the dominant action a bit more often than the other. The game thus features a trade-off in the way private information is used by Player 1, and how beneficial it is to hold such private information (the only exception is the last stage of the game, when the dominant action will always be chosen, because there is no longer any possibility to exploit the signal contained in your choices). This kind of game thus allows one to study the extent of the use of information, and the value of private information, i.e. how much more the informed player is able to earn.

The equilibrium strategies depend on two crucial features of the game: the length, denoted T, which is the number of stages during which both players play in the same matrix, and the structure of payoffs. The payoff structure we just described (the one shown in Figure 1.9.a) is called a non-revealing game, because the optimal strategy for Player 1 is to not reveal their private information in all stages but the last one: at equilibrium, it is best for Player 1 to behave as if the information were not available and the randomly drawn matrix were unknown. In the payoff structure shown in Figure 1.9.b, the prediction is exactly the opposite: the optimal strategy is for Player 1 to actually reveal private information about the true state of the world, by going straight for the stage game dominant action despite the loss incurred through sharing this information with Player 2. These theoretical predictions are summarised for different lengths of the game in Table 1.3. Is is worth noting that this change in the predictions is entirely due to the change in the payoffs. Before turning to empirical evidence on this game, you should try to think of each of the two matrix pairs, and ask yourself whether the way you will play the game will change so dramatically with the payoff structure.

A last theoretical prediction about this kind of game is that the expected payoff of Player 1 (known as the 'value of the game') is bounded above by the value of the infinitely repeated game (shown in the last column of the left-hand side of Table 1.3), and bounded below by the value of the average game. These theoretical predictions have

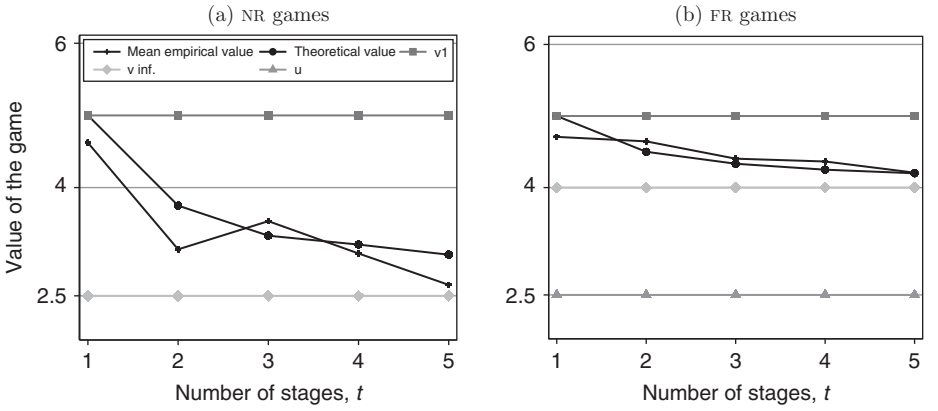


Figure 1.10 Empirical value functions

Note. Observed average payoff in the NR and FR games, along with the theoretical upper (v_1) and lower (∞) bounds.

Source: Jacquemet and Koessler (2013, p. 112, Figure 8).

been tested by Jacquemet and Koessler (2013) in an experiment in which participants play either the NR game or the FR game.

Figure 1.10 provides an overview of a comparison between the average observed values of the games in the experiment (measured as the average payoff earned by Player 1), and the predicted values as presented in Table 1.3. The empirical value functions confirm the theoretical bounds discussed above: the empirical value in both games lies between the value of the infinitely repeated game and the value of the average game. The empirical value is decreasing and smoothly converges towards its lower bound. This provides support for the theoretical analysis of the game. But the most challenging prediction is about the individual strategies, and their change according to the payoff structure.

Figure 1.11 provides information in that regard, through a summary of how information is used in each treatment. Remember that all the treatments have one prediction in common: Player 1, who knows which matrix has been drawn, has nothing to lose by using their private information (i.e. playing the stage-dominant action) at the last stage of the game. The figures are thus separated according to the stage within each game: the last stage of all games is reported on the left-hand side and the intermediate stages of all repeated games (in stages $t = 1$ to $t = T - 1$ for all $T > 1$) are reported on the right-hand side. From both the left-hand figure and the frequency of the stage-dominant action observed in the FR and NR games, experimental subjects unambiguously use information whenever it is worthwhile to do so. The relative frequency of the dominant action in the FR games is always higher than 90% and is much the same as in the last stage of the NR games. This frequency is much lower during intermediate stages of NR games, and is lower and lower as the overall duration of the game increases – when the revelation of information becomes more and more costly. Thus experimental subjects adjust their use of information not only as a reaction to experimental treatments, but also according to the decisions taken during the different stages of a given game.

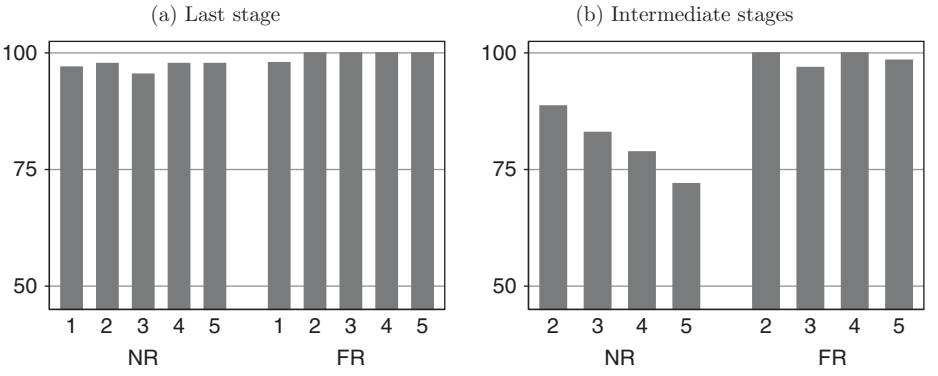


Figure 1.11 The actual use of information: informed players' behaviour

Note. For each treatment and each length, the figures display the mean share of the informed player's decisions that are the current stage-dominant action, in the final stage (*left-hand side*) and in intermediate stages (*right-hand side*).

Source: Jacquemet and Koessler (2013, p. 116, Figure 10).

Overall, empirical behaviour is relatively consistent with theoretical predictions in this environment, in sharp contrast with the two previous examples. This shows that complexity – in the game structure, but also in the theoretical predictions it induces – does not necessarily induce a larger gap between theory and empirical behaviour. The reasons for this consistency, in sharp contrast with the previous examples, is still a largely open question.

1.3.4 Beyond the Examples: Experimental Economics and Behaviour

These examples are not meant to provide a complete picture of the state of the art. But they do offer several important insights as to how experiments can help us better understand decision-makers. First, they show that experiments and economic theory are closely related. Empirical questions and the way data can be most usefully analysed are all based on a theoretical understanding of the situation. Second, and perhaps more importantly, the results described above shows a wide range of conclusions regarding the empirical relevance of theoretical results. Theory seems to accurately predict the outcomes in some games, and fails to do so in others based on similar behavioural assumptions. But the empirical relevance of theory goes beyond predicting outcomes. In particular, it accurately identifies the trade-offs and incentives people face, and how they are likely to resolve these issues. The above examples show that theory is often empirically influential in achieving this goal.

At the same time, it is also true that many behaviours and observed outcomes differ radically from theoretical expectations. Over the years, observations of this type have led specialists to enlarge the scope of the driving forces behind behaviour, to include psychological and sociological motives (this is the aim of behavioural economics). As the examples illustrate, the behaviour observed in economic experiments is related to theory in a complicated way: at times the *homo œconomicus* and human beings act as if

they were perfect strangers, and at other times they are surprisingly close to one another. How, why and under what circumstances do behavioural economics and economic theory converge or diverge? These are the core matters now being taken into consideration in the field (see, e.g. McFadden, 1999, for a survey).

1.4 Experimental Economics Today: What Every Newcomer Must Know

Since its tentative first steps, described at the beginning of the chapter, the use of experiments in economics has grown rapidly and dramatically. A very large number of contributions in economics nowadays rely on assumptions on individual decision-making, about which experiments definitely have something to say. What every newcomer must know in order to become familiar with experimental economics is so vast that no single work could possibly cover the whole field. This book is no exception. Instead, the following section offers an overview of the must-knows of experimental economics. Each of the items listed below corresponds to an index entry (see p. 441) that will refer the reader to sections of the book that discuss or illustrate this particular aspect. The section concludes the outline of the book, describing the must-knows this book will focus on.

1.4.1 Must-know 1: Microeconomic Theory and Decision Sciences

As explained above, experimental economics has grown together with game theory and decision theory. As a consequence, an important part of experimental economics focuses on assessing the empirical content of theories of behaviour. This requires familiarity with a vast number of topics from microeconomic theory. The most important of them are as follows.

- **Decision theory**. This strand of literature tries to better understand how individuals make decisions under risk and uncertainty, what role time-preferences and discounting of the future play and what leads to choice inconsistencies.
- **Game theory**. Agents in an economy interact with one another; their behaviour is directly influenced by the decisions of other agents and, in particular, by the beliefs they may hold about future behaviour of these other agents. Such considerations lead to strategic decision-making, which is a major topic in experimental economics.
- **Non-standard preferences**. The focus on the driving forces of individual behaviour led to challenging the standard way of looking at preferences. Alternative views of behaviour have been developed and are now part of the economists' toolbox. This includes non-standard decision models, such as prospect theory, where contingent states of the world influence decisions; and social and other-regarding preferences, according to which people's preferences not only are defined by consequences for themselves, but also account for the situations of others.
- **Aggregation**. Society has to make decisions, and thus needs to aggregate in one way or another individual tastes. This is the focus of auction theory, the analysis of markets and studies of collective decision-making, such as voting.

- **Psychology of behaviour**. The focus on individual decision-making makes it natural to borrow much from psychology. A large part of this literature is devoted to exploring the systematic deviations from rational decision-making, associated with several well-documented biases such as anchoring and status quo bias, endowment effect, confirmation bias, conjunction fallacy, framing effect, illusion of control, loss aversion.
- **Neuroeconomics**. The analysis of individual behaviour also borrowed in recent years from decision theory in medical sciences, leading to the field of neuroeconomics, which uses physiological measures to relate behaviour to its physiological driving forces.

1.4.2 Must-Know 2: Games and Decision-Making Frameworks

The implementation in a laboratory of the theoretical frameworks described above often makes use of environments, procedures and rules of particular types. They are tools designed to study different aspects of individual behaviour. They are nowadays considered part of the standard toolbox of anyone working in the field.

- **Elicitation procedures** are mechanisms that force agents to reveal something about themselves, such as risk or intertemporal preferences, or beliefs about what others will do.
- **Experimental games** are games structured with specific theoretical properties that are widely used and studied in experimental economics. These key games include the prisoners' dilemma, the trust game, the stag hunt game, the dictator game, the guessing game, the ultimatum bargaining game, the voluntary-contribution mechanism, the minimum effort game and many others.
- **Psychological questionnaires** can be used to gather data on how people think through their decisions and how they consider different situations. Psychometric questionnaires include, for instance, measures of cognitive and non-cognitive skills, personality traits or emotions.

1.4.3 Must-Know 3: Fields and Applications

The insights from (micro)economics that are implemented in the laboratory can be applied to a wide range of field applications. As a result, there is a growing literature of experiments contributing to a better understanding of issues related to the various fields of interest to economics. Among them, the most important are:

- **Labour economics**, which focuses on the effects of labour market policies, the trade-off between consumption and leisure, the education production function, etc.
- **Personel economics** focuses on how people behave in firms, dealing with questions such as how people choose jobs and the reasons why they choose these jobs, how much people work and how they respond to monetary and non-monetary incentives.
- **Industrial organisation** focuses on how firms interact with one another under decision variables of different kinds, such as volumes, prices or levels of advertisement,

and different market structures, such as auctions, oligopoly or perfectly competitive markets.

- **Environmental economics** studies the policies designed to discipline behaviours that are detrimental to the environment, dealing with problems such as greenhouse gas emissions, air pollution, water quality, toxic waste or global warming. The issue of collective decision-making and the problem of free-riding are of critical importance in this domain.
- **Health economics** is a field concerned with the health of individuals, the health care market, the supply of health services or the public health system in general. Preventive health care is an important behavioural issue, for example, and the supply of health care services by physicians raises intriguing questions about incentive design and payment schemes.
- **Law and economics** tries to understand how individuals react to different sets of legal rules. The focus is on circumstances that make people comply with the law, and how the law changes social norms and equilibria.

1.4.4 Must Know 4: Methodological Issues: Outline of the Book

Lastly, laboratory experiments are a very precise way of gathering data and providing an empirical counterfactual to microeconomic theory. This comes with drawbacks and advantages, with several constraints on how experiments are run, and with questions regarding what they tell us about relevant economic issues. The aim of this book is to provide a review of the current answers to this strand of questions.

Chapter 2 is an introduction to the field, by describing step by step what an experiment looks like from the point of view of a participant, before turning to the analysis of the same experiment. This is a critical starting phase in becoming an experimenter, as running experiments is all about understanding how people behave, and avoiding any misunderstanding they may have about the environment. The best way to deal with this issue is to imagine how you would act if you were a participant in an experiment. The second important lesson from this introductory chapter is that experiments involve many unusual procedures and implementation rules, which may not appear quite appealing at first glance. The last part of the chapter describes the reasons why each of these features is required to make an experiment convincing – and will discuss what *convincing* means for an experiment.

This book is divided into four parts. As is explained above, Part I provides an overview of what experiments are. Part II explains why experiments in economics are needed and to what extent they are useful for empirical research in economic science. Each chapter provides a specific answer to this question. In Chapter 3, we describe how experimental economics is related to other empirical methods in economics. Basically, experiments provide a way to choose the data-generating process, enhancing the ability to measure unknown quantities relevant to economic analysis. Chapter 4 turns to the relationships between experiments and economic theory. We will see that experiments serve three different purposes: testing theory in a controlled environment, searching for facts and

whispering in the ears of princes. Theory and experiments share a dynamic of mutually informing each other in this process.

Part III describes how laboratory experiments can achieve these goals. Each chapter explains how to produce experimental results, one step after the other. Chapter 5 focuses on how to design an experiment such that observed behaviour can be related to the institutions under study – i.e. which is internally valid. Chapter 6 covers all the practical aspects required for running an experiment. These practical aspects include all the phases, from building a laboratory well upstream to the final laboratory session. Last, in Chapter 7, we review the main statistical methods that are commonly used to analyse experimental data.

The focus of Part IV is to assess the relevance of what laboratory experiments tell us. Each of the chapters presents an overview of areas in which experimental results are able to shed additional light on existing knowledge. Chapter 8 begins with a question called the 'external validity' of experiments: what do decisions taken in the artificial framework of a laboratory tell us about real life? When an experiment satisfies the conditions so as to be both internally and externally valid, then the experimental results can be used by economic theory and public policy. This opens the way to a more general discussion on the possibility of inductive reasoning in economics, an issue covered in the first section of Chapter 9. This discussion will also show that observed behaviour in the lab has drastically changed the way economists think of institutions and how to organise collective decisions. This point will be the focus of the last sections of this chapter, on the design of public policies thanks to the lessons drawn from the laboratory.

Summary

This introductory chapter presented the field of experimental economics from a general perspective. Originally, experiments in the social sciences and in economics, in particular, were thought to be impossible. The first experiments beginning in the second half of the twentieth century showed otherwise. However, experimental economics did not truly break through until the focus of economics changed with the fall of general equilibrium as the central theory and the questions started to turn more towards issues related to human behaviour. To illustrate the current state of the art, we reviewed three examples from the experimental literature testing behavioural insights from game theory: the prisoners' dilemma, the centipede game and a repeated zero-sum game with incomplete information. Observed behaviour and theoretical predictions may not match up perfectly but they are not perfect strangers to each other either. This summarises the current state of the art in the field: the core issue at stake in ongoing experimental research is identifying situations where theory goes wrong and where it performs well. Since experimental economics has developed together with the use of decision and game theory in economics, the range of topics to which experiments are applied is now far too wide to be reviewed in a single book. This book focuses on experiments as an empirical methodology to inform economic science.

2 A Laboratory Experiment: Overview

The goal of this chapter is to introduce the methods used in experimental economics to study people's behaviour. This book will subsequently focus on how to design laboratory experiments, how they allow measurement of interesting and relevant parameters and how to interpret the empirical conclusions drawn from the experiments. Chapter 6, in particular, will describe the practicalities related to the implementation of laboratory experiments. Before getting to this material, we would like you, the reader, to learn what a laboratory experiment looks like from the inside. To that end, you will see things from the point of view of what we will call a *subject* or *participant*: you will be a person who comes to the laboratory to be involved in an 'experiment in economics, hence contributing to scientific researches' (this is more or less the kind of general statement one finds in the advertisements used by experimental laboratories across the world to recruit subjects).[1]

To be sure anyone reading this book will actually go through this preliminary step, let us add a few words about why we believe that to truly learn about laboratory experiments it is essential to do so from the inside at least once in your life (it goes without saying that such an experience will be best achieved by being involved in an actual experiment in a department close to your location, if such an option is available). The first reason is that, as an economist, you will certainly have your doubts about the value of the results generated by laboratory experiments. A great many experimental results seem either wonderful, trivial or silly at first glance. Once you have made the effort to mentally represent how you would have behaved in a given situation, we think you will be well prepared to better understand and use experimental economics' method and results.

A second, even more important, reason is that a good part of this book will be devoted to explaining how to carefully design laboratory experiments. The term *carefully* will mean *in such a way that observed behaviour delivers general lessons about the properties of the decision-making environment*. The first thing you should learn about laboratory experiments is that observed behaviour comes from real human beings to whom you are describing the environment. Consequently, you, as an experimenter, are the one responsible for everything participants get, everything they miss and everything they misunderstand or get confused about. An important skill in order to achieve this goal is your ability to put yourself in a participant's shoes.

[1] The remaining describes the proceedings of a typical experiment based on the ones we know best, which is how they are run in our own departments. There are, of course, many location-specific variations: our aim is not to describe best practices, but to provide one detailed example of how experiments are run.

Lastly, as you will see in the paragraphs ahead, the methodology of laboratory experiments means that subjects will be involved in what they may find to be some very surprising procedures. The reasoning behind these procedures is not always obvious to the participants. Again, you may understand the issue more clearly if you are able to remember how you yourself felt as a participant. In this chapter you will be exposed for the first time, without any prior knowledge about the method, to this strange sequence of events called an experiment (the rationale of which will be described in Chapter 5).

Now try to imagine that you have signed up on a website to participate in an experiment. Soon after you register, you will receive an e-mail asking you if you would like to come on *HOUR, DAY-MONTH-YEAR* for an experimental session to take place at *ADDRESS*. You agree to participate and confirm your participation in the given experimental session.

2.1 The Experiment

When you arrive at the building, you will be welcomed by someone with the full list of people who have confirmed. You will then be asked to show an ID and once your identity is confirmed you will be given a form similar to the one shown in Figure 2.1. You will have to sign the form in order to participate in the experiment. Once the appointed time is reached, all the people waiting in the hall will be asked to go through the university building to the door of a lab room.

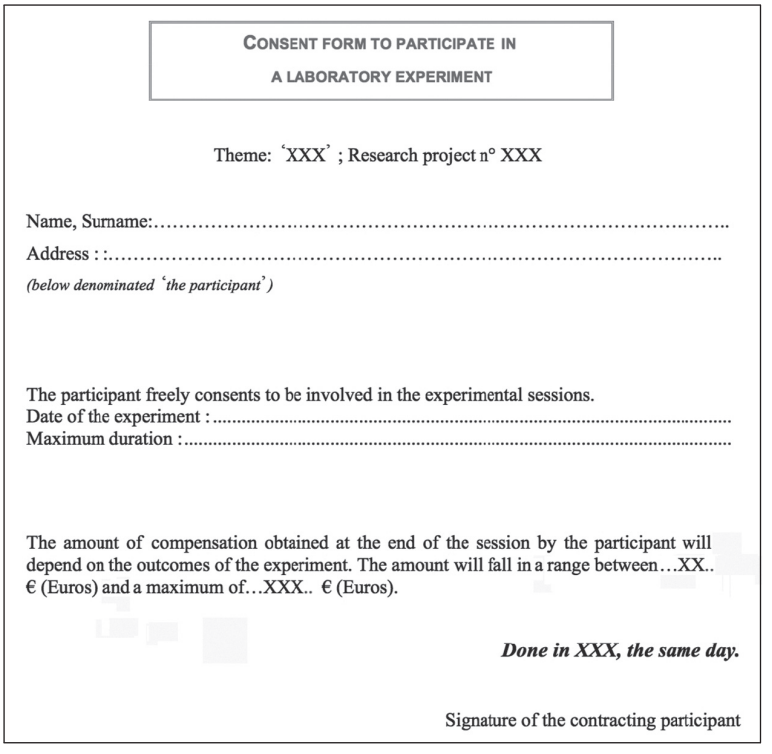
In front of that door, the same person who has welcomed you will explain the following:

> You will enter one by one into the room behind me, which is the laboratory where I will explain everything you need to know to participate in this experiment. Before going in, I will ask for the consent form you signed, and have you take a sheet of paper. The name written on this sheet is the name or number of your computer, where you will sit once you are in the room. Once everybody has entered, we will all start the experiment together as a group. Meanwhile, please wait quietly; thank you for your patience.

Before you begin the experiment, you will be informed about the way it will proceed. To that end, you are given a sheet of paper with the following text, the *instructions for the experiment*. The experimenter reads it aloud and encourages you to carefully follow on your own paper – that is yours for the entire duration of the experiment. Let's read it together.

Instructions for the experiment

You are participating in an experiment in which you can earn money. The amount you earn will depend on your own decisions as well as the decisions of the other participants. Before starting the experiment, we will ask you to answer a few questions in order to get to know you better (your age, gender, occupation, etc.). **All this information will remain anonymous and confidential.**



CONSENT FORM TO PARTICIPATE IN
A LABORATORY EXPERIMENT

Theme: 'XXX' ; Research project n° XXX

Name, Surname:……..

Address : :……..

(below denominated 'the participant')

The participant freely consents to be involved in the experimental sessions.
Date of the experiment : .....
Maximum duration : .....

The amount of compensation obtained at the end of the session by the participant will depend on the outcomes of the experiment. The amount will fall in a range between…XX.. € (Euros) and a maximum of…XXX.. € (Euros).

Done in XXX, the same day.

Signature of the contracting participant

Figure 2.1 Consent form

Procedures for the experiment

At the beginning of the experiment, **two groups, each involving 9 participants**, will be formed. **Each participant belongs to the same group during the whole experiment**.

Overview. You will be participating in an auction in which you are the buyer. The currency unit used in the auction is the ECU (Experimental Currency Unit). Its value in euros is described at the end of the instructions. You will submit a bid in ECU to acquire one unit of the good which the experiment monitor then will reacquire from you. There will be several rounds of bidding. The outcome of the auction in each round directly influences how much you will be paid at the end of the experiment.

Procedures for each round

Each round has 8 steps.

Step 1. Each bidder looks at his or her **resale value** on their screen. We label **resale value** the price in ECU that the monitor will pay to buy back a unit of the good that is purchased in the auction. **The resale values of different participants in a group can be different**. Once you have looked at your resale value, press the OK button.

Step 2. Each bidder then submits a bid in ECU to buy one unit of the good. To do this, move the scroll bar up or down until you see the price you want to submit. Then press the OK button below the scroll bar to confirm your choice.

Step 3. The monitor will rank the bids from highest to lowest. For instance:

n^o 1 fs.l ECU **Highest bid**
n^o 2 df.g ECU
n^o 3 za.f ECU
n^o 4 sc.d ECU
n^o 5 qs.a ECU
n^o 6 nj.h ECU
n^o 7 hh.m ECU
n^o 8 ht.t ECU
n^o 9 ky.l ECU **Lowest bid**

Step 4. The second-highest bid (bid n^{o}2) determines the **market price**. In the above example, if the second-highest bid is df.g ECU then the market price would be df.g ECU:

n^o 1 fs.l ECU

n^o 2 df.g ECU **Second-highest bid: market price**

n^o 3 za.f ECU
n^o 4 sc.d ECU
n^o 5 qs.a ECU
n^o 6 nj.h ECU
n^o 7 hh.m ECU
n^o 8 ht.t ECU
n^o 9 ky.l ECU

Step 5. The buyer who bids the highest price (the buyer ranked n^{o}1) purchases one unit of the good at the market price. In the above example the buyer who bid fs.l ECU purchases one unit of the good that costs df.g ECU.

Step 6. Buyer n^{o}1 then sells the unit back to the monitor. The price of this transaction is the resale value listed on the screen for that round. The profit in ECU that bidder n^{o}1 earns for that round is the difference between the resale value and the market price:

$$\text{profit} = \text{resale value} - \text{market price}$$

Important remark. Your profits can also be negative: if you buy a unit of the good and the resale value is less than the market price, your profits will be negative.

Step 7. All bidders at or below the market price (buyers n^{o}2 to n^{o}9) do not buy anything, **so they make zero profit for that round**.

Step 8. End of the round. Your profit in ECU in that round appears on your screen. Press the OK button once you have noted it. Your screen will then indicate whether a new round is about to begin or the experiment is over.

How will you take your decisions?

Your screen is divided into three areas:

All the information you need to take your decisions will be displayed in the upper part of the screen.

You then take your decisions by pressing on the buttons displayed in the middle part of the screen.

The bottom part will show you your past decisions and profits.

Payment of your earnings

At the end of the experiment, we will compute the sum of your profits in ECU across rounds. If your profit in a given round is negative, the total decreases; if your profit in a given round is positive, the total increases. This total is converted into euros according to the rate 3 ECU = 1€. A fixed fee equal to **10** € is added to this payoff. You will be paid the corresponding monetary payoff in cash privately at the end of the experiment.

Please do not talk and try not to communicate with any other subject during the experiment. If you communicate, you will be asked to leave and forfeit any money earned. It is essential that you understand the instructions correctly. If you have any questions, please raise your hand and someone will come and answer them. Please be sure to follow these instructions.

Thank you for participating.

If you have any questions about the instructions, you will be able to quietly raise your hand. The experimenter will then come over to you to answer your question(s) in private. Obviously, in this textbook experiment you cannot raise your hand to ask questions. So please take the time you need to read and reread the above text, until it is perfectly clear in your mind how the experiment will be conducted and what happens depending on what you and the others do . . .

Good. Now let's go through the last stage: a short questionnaire to check that everything in the instructions is crystal clear for you. The answers provided on the questionnaire will not influence your earnings or participation in the experiment in any way: its only aim is to help you be sure everything is clear to you. Please fill in the questionnaire.

Pre-experiment questionnaire

1. New groups are formed after each round.

   ☐ YES ☐ NO

2. Each group includes _____ participants.

3. At the beginning of each round, all the participants in my group are attributed the same resale value.

   ☐ YES ☐ NO

4. When I make a bid, I can bid any amount I want.

   ☐ YES ☐ NO

5. The market price is set by the bid of the second-highest bidder in my group.

   ☐ YES ☐ NO

6. If my bid is the highest bid and is equal to rr.u ECU and the second-highest bid in my group is gg.k ECU, then I buy the unit of the good.

   ☐ YES ☐ NO

   If YES, I pay: _____ for the good.

7. If I purchase a unit of the good and my resale value is greater than the market price, I will make positive profits.

☐ YES ☐ NO

8. The monetary payoff I will be paid at the end of the experiment depends on the amount of ECU I earned during the auction.

☐ YES ☐ NO

Now that you are done, let's go through the questionnaire together, and stress the answers that accurately describe how the experiment will proceed.

Q1 '*New groups are formed after each round*'. The answer is **No**. The groups remain the same during the entire experiment.

Q2 '*Each group includes **9** bidders*'.

Q3 '*At the beginning of each round all the participants in my group are attributed the same resale value*'. The answer is **No**. The resale value of the various participants in your group can be different.

Q4 '*When I make a bid, I can bid any amount I want*'. The answer is **Yes**. There is no constraint imposed on the price you choose.

Q5 '*The market price is set by the bid of the second-highest bidder in my group*'. The answer is **Yes**, the market price is the second-highest one among all bids chosen in your group.

Q6 '*If my bid is the highest bid and is equal to rr.u* ECU *and the second-highest bid in my group is gg.k* ECU*, then, I buy the unit of the good and I pay **gg.k**,* i.e. *the market price.*' The answer is **Yes**.

Q7 '*If I purchase a unit of the good and my resale value is greater than the market price, I will make positive profits*'. The answer is **Yes**, since your profit will be computed as your resale value minus the market price.

Q8 '*The monetary payoff I will be paid at the end of the experiment depends on the amount of* ECU *I earned during the auction.*' The answer is **Yes**.

If you are surprised by any of these answers, please read again the instructions sheet carefully before returning to the question you had doubts about.

Now that all the written material has been read, the experiment can start. The first step is illustrated in the screen capture in Figure 2.2. The display has three frames. The top frame gives you information about the current round. The middle frame is devoted to your own decisions. The bottom frame provides a reminder about previous rounds. The information provided here includes the round number and three key elements: your resale price, the price you have chosen and your profit in each round. The first display, on the top of the screen, gives you a resale value for the good (24 ECU). All you have to do is to click on the OK button to proceed.

Once you have clicked on the OK button, a second screen appears as shown in Figure 2.3. This second display asks you to choose a price for the good. To choose the price

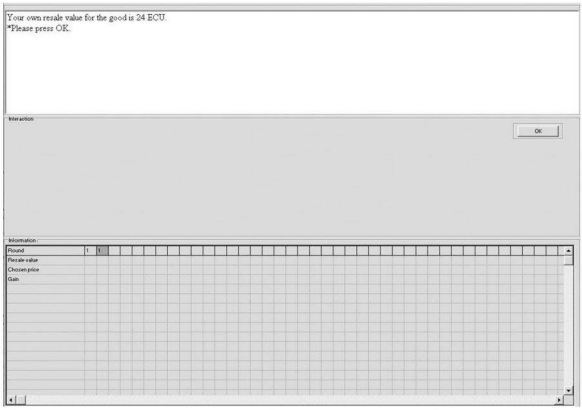


Figure 2.2 First screen: resale value in the first round

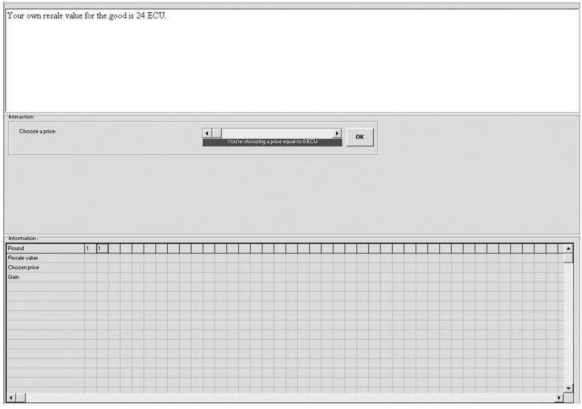


Figure 2.3 Second screen: bid in the first round of play

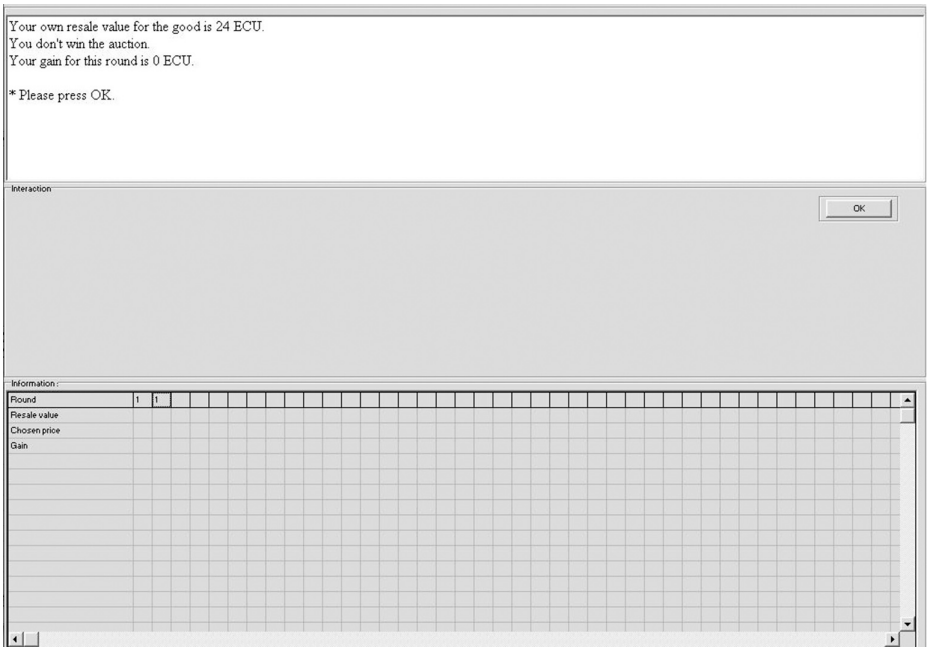


Figure 2.4 Third screen: results of the first round

you are willing to bid you have access to a scrollbar in the middle frame. Once you have chosen the price you want to bid, click on OK to proceed. Let us say, for the sake of the example, that you chose a price of 37 ECU for that round. The third display in Figure 2.4 shows you the result of the auction. The information about whether or not you won the auction appears on your screen, saying you didn't win the auction. Your resale value for the good and the profit you gain from that round (0 here) is also reported in the top frame. Click on OK to proceed and a second round will start.

Let us skip the subsequent rounds, and jump to round 6. The display you get at this stage is shown in Figure 2.5. The bottom frame shows the past experiences you have had with the auction. In round 1, your resale price was 24 ECU, your chosen price was 37 ECU and your profit was 0 ECU. In round 2, your resale value was 84 ECU, your chosen price was 86 and your profit was equal to 8 ECU. From this you can infer that the second price was equal to 76 ECU. In rounds 3 to 5, you gain nothing from the auctions. Now, in round 6, the top frame assigns you a resale value for the good equal to 65 ECU. As before, you have access to a scrollbar in the middle frame to choose the price you want to bid. Once you have chosen the price you are willing to bid, click on OK to proceed. Let us state that you are willing to bid 88 ECU for the good. As shown in Figure 2.5, your bid was high enough to win the auction and the profit you made is equal to –19 ECU. On the screen, you can also see the second price you paid for the purchased good. This price is equal to $65 + 19 = 84$ ECU. Click on OK to proceed and a new round will start.

Once you have finished with all the scheduled rounds, the screen shows a summary of all rounds and the resulting monetary gain you earn from the experiment. All participants

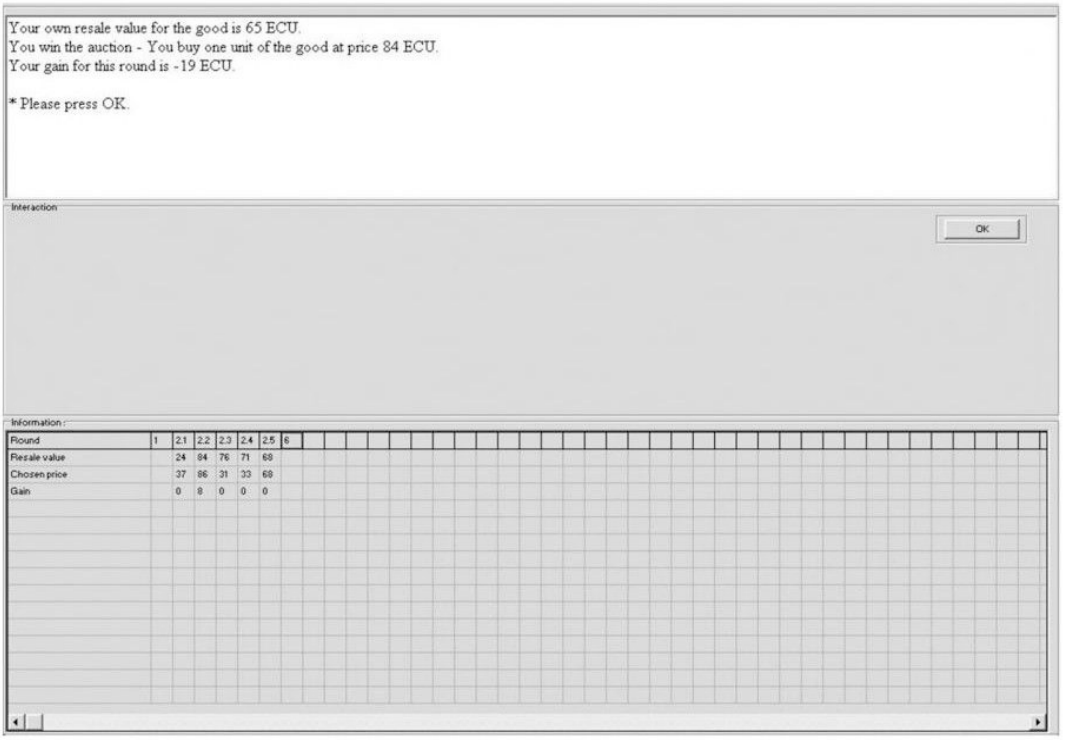


Figure 2.5 The sixth round of the experiment: screen captures

come one by one to a separate room inside the lab, and get paid in cash according to this amount. You are then thanked for your participation, and you can leave the room where the experiment was run.

2.2 The Experimenter's Role: The Game under Study

If you belong to the same population as the students that usually come to experimental economics laboratories, then you probably did not recognise the rules of the mechanism described above. Otherwise, you probably have already guessed that these instructions are meant to put subjects into a second-price auction.

Before getting to a description of the experiment from the experimenter's point of view – what is called *the design of the experiment* or *experimental design* – you need to become familiar with the questions answered thanks to experiments of this type in empirical research in economics. These answers come from the extremely powerful properties of such mechanisms according to economic theory, as shown in detail in the next section.

2.2.1 Theoretical Properties of Second-Price Auctions with Private Values

The above experiment considered a market with one single seller and n buyers. Only one unit of the good was to be sold on that market. Here, we are not concerned by the seller's

behaviour, so we can assume the good is worth 0. In that case, the seller agrees to sell the good at any positive price.[2] The main aim of economic theory in such a context is to find the equilibrium. The equilibrium here signifies who buys the good and what price is paid for it, given economic agents' preferences.

Let us start by formalising the buyer's preferences. We denote v_i the monetary equivalent of the buyer's i utility of obtaining the good, aka the *private value* of the good. The value is private because different people, i, are allowed to have different values. Moreover, an individual i does not know the others' value (i.e. values v_j for $j \neq i$). This utility of owning the good is a measure of the buyer's willingness to pay because if the buyer trades a monetary loss equal to v_i against the consumption good, their situation remains exactly the same. If q is the price to be paid for consuming the good, the buyer's benefit is thus exactly equal to $v_i - q$. The seller's profit is equal to $\pi = q - 0 = q$. Thus, in moving from the seller's hand to the buyer's hand, the good generates a value equal to v_i. The higher the buyer's private value, the higher the amount of wealth created in the way the available scarce resource (one unit of the good) is allocated. This observation highlights the first important property of an allocation mechanism: its efficiency. Efficiency is measured here as the ability of the allocation mechanism to allocate the good to the agent who attributes the highest value to it, to achieve the highest possible level of wealth.[3]

An examination of the payoff functions of the two agents clearly shows that the price is not relevant to efficiency. A change in price is a zero-sum transfer between the buyer and the seller, leaving unchanged the total amount of wealth. The price only decides on how the surplus is shared between the two economic agents. However, buyers will not agree to pay just any price, because they would incur a loss were they to pay a higher price than their private value v_i to acquire the good. This means that the price the buyer announces in the auction (what we called a *bid*) is correlated with their privately known valuation v_i. As a result, the bid becomes an observable signal of the buyer's own preferences. This points to the second important property of an allocation mechanism: its revelation property, measured as the informativeness of the bid regarding the underlying true preference.

How a given allocation mechanism performs on both these levels depends on the bidding behaviour it induces. The second-price auction implements the following allocation rule: the buyer with the highest bid wins the auction, but the price paid for acquiring the good is the second-highest bid. In order to elicit what a given buyer i can best do in such a context, let us denote their bid b_i and B_i the highest bid chosen from among the other $n - 1$ bidders.[4] Since values are private, buyer i knows nothing about B_i. What they know for sure, though, is that they will win the auction by choosing a bid b_i higher than

[2] One possible rationale for this assumption is that the good is already produced and has no consumption value for the seller.

[3] An alternative view is that efficiency also exhausts all possibilities of trade: if the mechanism allocates the good to an agent whose value is lower than that of another, then the two could find a mutually beneficial agreement to trade the good – since another agent in the economy is willing to pay more for the good than what it is worth for its current owner. This applies for as long as the final allocation remains sub-efficient.

[4] Formally, $B_i = max_{j \neq i} b_j$, which can also be denoted $B_i = maxb_{-i}$, with $-i$ denoting all the bidders except i.

B_i, whatever B_i. And that, in this case, B_i will be the price they will pay for the good. Thus, buyer i (weakly) prefers to win the auction if $v_i \geq B_i$ (so that $v_i - B_i \geq 0$) and prefers to lose it otherwise (if $v_i - B_i < 0$). Consider each possible state of the world in turn.

If, on the one hand, B_i is strictly higher than v_i, then buyer i prefers to lose the auction – because winning it would mean a decrease in wealth. There is one choice which guarantees that the buyer is sure to lose the auction in all instances such that $v_i - B_i < 0$: the bid b_i simply should never be higher than the private value v_i. If, on the other hand, B_i is lower than v_i, buyer i will want to win the auction. You might have already noted that the outcome of the auction stays the same whatever the bid b_i is, as long as it is higher than B_i. In a second-price auction, the price q to be paid will always equal B_i and buyer i will remain the winner. For any bid b_i chosen in such circumstances, moving to a higher bid increases the probability of winning the auction without changing the price to be paid, provided that such a bid remains compatible with this particular state of the world. The threshold separating the two states of the world is reached when the bid b_i reaches the private value v_i. At that point, the chances of winning the auction against a B_i higher than v_i become positive.

As a result, the optimal bidding strategy for buyer i in this environment is to choose a bid b_i exactly equal to their private valuation $b_i^* = v_i$ as it provides the highest possible likelihood of winning the auction when and only if it is desirable. This strategy is followed by all bidders on the market, leading to the two main theoretical properties of the second-price auction. First, the ranking of equilibrium bids is exactly the same as the ranking of valuations: the winner of the auction will thus be the one with the highest private value, so that the mechanism achieves an optimal allocation. Second, the equilibrium bidding behaviour induced by the auction is *perfectly revealing*: each bid is a perfect signal of the buyer's underlying true preference. For daily-life goods, either market ones like a pen or a coffee or non-market ones like those described in Focus 2.1, these preferences are privately known by their holder, and unobservable. In such contexts, the auction can be used as a revelation mechanism – a way to better know what the distribution of individual preferences is.[5]

2.2.2 Why Experimental Auctions Are Important

The above-mentioned properties of second-price auctions are all the more impressive given the observational context of the model's assumption. While each buyer is the only

[5] This last property is the main difference from the behaviour induced by the first-price auction mechanism. In first-price auctions the winner's bid is also the market price. There, the bid determines both who wins and the price the winner will pay. A unique instrument is used to select the winner and to offer the winner's prize. The second-price auction, by contrast, relies on two instruments each serving only one objective: the winner's bid determines the winner while the second-highest bid determines the profits. The optimal strategy in first-price auctions thus resolves the trade-off between winning the auction on the one hand, and maximising profit in the event of winning on the other hand. This results in an equilibrium bid slightly lower than the private value, so that the mechanism is not perfectly revealing. But the allocation achieved is still efficient, as all bidders will adjust their strategy in this way and the ranking of bids will perfectly match the ranking of private values.

Focus 2.1
Preference elicitation and policy-making: the hypothetical bias

Most non-market goods – such as environmental protection, new medicines or amenities like a park in a city – are publicly funded. When it comes to deciding on how much to invest in one particular good of this type, it is necessary to run a cost–benefit analysis. In particular, decision-makers have to assess the benefits of the investment. One key aspect of these benefits is how desirable the good is for the target population. The desirability of the good is central not only because it is a measure of how much welfare the investment generates but also because it measures the willingness to pay for the good of the people who will consume it. Obtaining such a measure amounts to eliciting people's true – unobservable – preferences for the good. To that end, the most commonly used method is to conduct surveys requesting respondents to state how much they would like to pay to have the good produced. But a widely recognised challenge of such an elicitation method is that people may declare a higher amount than they are actually willing to pay for the good, because the question is asked in a hypothetical context without real economic commitments or consequences (Arrow et al., 1993). Of course, relying on such stated preferences if hypothetical bias does exist leads to many inefficient investment decisions. Revelation mechanisms, such as the second-price auction, are thus studied experimentally in order both to study the phenomenon, and to find settings that overcome the bias.

one to know their own individual preferences, and knows only that, one merely has to ask each buyer for a simple bid. Beforehand, though, each buyer has to be warned that the allocation of the good will follow the rules of a second-price auction. Based on these very simple market rules, the outcome of the mechanism is to allocate the good to the buyer everyone would have chosen based on a perfect knowledge of the full demand function (in the aim of achieving the highest possible level of wealth from the available resources). Moreover, an inspection of the individual bids even provides perfect knowledge of the underlying individual preferences. Focus 2.1 describes the consequences of this property for policy design. The remarkable properties of second-price auctions remain theoretical, however. In other words, all of this is conditional on the empirical relevance of the bidding behaviour predicted by the model.

These are the kinds of question that laboratory experiments are particularly good at answering. This is so for two reasons. First, the experimental context provides more control over the setting in which behaviour is analysed. Second, an experimental context provides more measures of what happens in different circumstances. We will detail these two reasons in turn.

First, note that the experiment we described in Section 2.1 is the exact empirical counterpart of the model. The allocation rule is the same, but even the payoff functions driving individual preferences parallel those embedded in the model. The resale values v_i, in particular, play the same role in the empirical world provided by the experiment as the private values do in the world described by the model. This means that the experimenter 'chooses' or induces the participants' preferences; for that reason, this kind of setting is called an *induced value* experiment – as the experimenter chooses,

induces, the individual preferences according to which people in the experiment are making decisions (see Section 5.3.1).

At the same time as they are induced – hence controlled – individual preferences are also perfectly observed by the experimenter. This represents a huge added value compared to what is generally available when one is working on data from the real world. Such an observation is crucial as theoretical predictions are all about the relationships between the outcome of the auction (the list of bids, the identity of the winner, the market price, the winner's profit) and the underlying factors driving such outcomes, namely preferences. As a matter of fact, the experimental context allows one to observe not only the market price but also the complete list of all the prices proposed for each bidder and each auction. Again, this is not the case in most observational data available from real-world auctions (where one typically observes only the winning bid). These features make the empirical outcomes from experiments highly reliable, as will be explained in Chapter 3.

As a result of this combination of strong control capabilities and wide observation opportunities, the experiment in which you have just participated provides much information on the theory of second-price auctions. First, one can test the efficiency of the allocation rule by comparing the private value of the winner to the highest private value present in the market. If the winner is the participant with the highest value in the market, then the allocation rule was empirically efficient. Second, the revelation properties of the mechanism can be empirically assessed. This assessment is based on the comparison between individuals' bids b_i and their private values v_i. If the bids correspond to the private values then the second-price auction is an accurate preference revelation mechanism. It is worth noting, on a final note, that theory serves as a benchmark for the empirical observations delivered by the experiment – the way data are analysed is framed and driven by the theoretical understanding of the environment. Chapter 4 will feature a more detailed discussion of this important aspect.

2.3 Experimental Second-Price Auction with Private Values

It is now time to move from the front stage to the backstage of the experiment and look at it from the point of view of the researchers rather than the participants. There are always some features of an experiment that are not visible to participants (hence not described in the instructions presented in Section 2.1) but are very important to know in order to fully understand the results. In the case of a second-price Vickrey auction, a significant part of the design lies in the way the resale values (the experimental equivalent of the private values v_i) are chosen. In the example opening the chapter, the second-price auction experiment involves nine bidders and nine bidding rounds. For the nine bidders, the list of induced values is $\{84; 76; 71; 68; 65; 63; 53; 38; 24\}$ and is kept constant during the experiment. From the point of view of the experimenter, $\{84; 76; 71; 68; 65; 63; 53; 38; 24\}$ is the induced demand curve. Each value in the list is allocated to exactly one bidder in each round. From one round to the next, each bidder's value changes as they pick up another

Table 2.1 Empirical revelation properties of a second-price auction

Induced value	24	38	53	63	65	68	71	76	84	All
Aggregate demand (AD)										
($v_i \times 18$)	432	684	954	1134	1170	1224	1278	1368	1512	9756
Revealed AD	492	678	816	1145	1121	1229	1260	1406	1490	9637
in % of AD	114	99	85	101	96	100	99	103	98	99
Round	1	2	3	4	5	6	7	8	9	
Aggregate demand (AD)										
($2\sum_i v_i$)	1084	1084	1084	1084	1084	1084	1084	1084	1084	9756
Revealed AD	895	1045	1141	1065	1174	1143	1116	1045	1013	9637
in% of AD	83	96	105	98	108	105	103	96	93	99

Note. The table reports the revealed demand in the experiment, i.e. the sum of bids posted by all bidders defined in columns. The upper part groups bidders according to their induced value; the lower part provides round-by-round data.
Source: Jacquemet et al. (2009, p. 38, Table 1).

one in the list which they have not yet tested. Since there are exactly nine bidding rounds (note that the instructions only say that there will be several rounds; the reason for this choice is to avoid *end-game effects*), such a rotating bidders-value matching puts into practice all possible permutations of the constant aggregate demand curve.

Table 2.1 shows the baseline results obtained by Jacquemet et al. (2009).[6] The behaviour displayed in the table stems from 18 subjects, resulting in two markets of nine subjects each in the experiment. Each subject participated in one of the two nine-person markets, with nine repetitions, each with different induced values. The experiment provides 2 (markets) $\times$ 9 (subjects) $\times$ 9 (repetitions) $=$ 162 observations of bidding behaviour (which are not independent, however, as is often the case with experimental data: Chapter 7 describes the statistical tools used to address such specificities). Table 2.1 is made up of two main parts. The upper part classifies the answers to each of the induced values proposed to the participants (remember that each induced value is proposed to each subject once, and only once, at some point of the nine repetitions of the second-price auction). The lower part classifies the answers with respect to the rounds in which the bids were revealed. Before we turn to commenting on observed behaviour of the participants in the experiment, try to recall how you decided to behave in the experiment and, in particular, whether any of the theoretical insights of Section 2.2 came to mind when you decided on a bidding strategy ...

In the upper part of Table 2.1, the second row reports the corresponding induced aggregate demand (*AD*) for each induced value in column (24, 38 ...). For each induced

[6] This design was first implemented by Cherry et al. (2004). The primary focus of Jacquemet et al. (2009) is to assess the influence of the origin of the experimental endowment on bidding behaviour. To that end, a preliminary step is added to the design presented in Section 2.1. Before participating in the auction, subjects answer a 20-item quiz in which the right answers are remunerated. These subjects thus enter the auction with an amount of money they consider as their own (because it was 'earned') rather than 'windfall'.

value, the aggregate demand is equal to 18 times the induced value. For example, when the induced value is equal to 24, the aggregate demand for the 18 subjects is equal to $24 \times 18 = 432$. As each induced value is allocated to exactly one subject in each auction round, the aggregate demand for a given induced value pools together participants irrespective of the order in which they obtained that value. For example, when the induced value was equal to 24, the aggregate demand pools subjects who acquired this value in first, second, ..., *n*th position. A consequence of this is that aggregate demand smooths out the effect of learning or experience (see Section 5.5 for a detailed discussion).

What subjects actually did is displayed in the third row of the upper part. The empirical counterpart of the aggregate demand, called revealed aggregate demand (*revealed AD*), is shown here. The revealed aggregate demand is the sum of the bids chosen by all the subjects who have experienced the induced values provided in a given column. For example, the sum of the bids of subjects with an induced value of 24 is equal to $2 \times \sum_{i|v_i=24} b_i = 492$. An empirically perfectly revealing auction should equalise aggregate demand and revealed aggregate demand. To make the comparison easier, the fourth row presents the ratio between the revealed aggregate demand and the aggregate demand. This ratio is called the revelation ratio, i.e. the share of the induced demand which is stated through the bids. For (almost) all induced values the average revelation rate is remarkably close to 100%; and it is true all along the demand curve, i.e. along the whole distribution of induced values. These results suggest that the second-price auction performs remarkably well in terms of demand revelation.

The picture is quite different in terms of the second theoretical property of the mechanism, namely its ability to implement an efficient allocation. Strictly speaking, efficiency can be fully assessed from the induced value of the winner in each round. In this experiment, the winner should always be the bidder with the maximum induced value, which is equal to 84. This actually happened in 61.1% of all cases, which is far from what theory predicts. The experiment also allows measurement of the extent of the loss associated with such sub-efficient outcomes: this can be measured by the share of the potential wealth that is actually realised in the experiment. The data reveal that the average induced value of the winner over the 18 auctions is 77.5 in the experiment, so that 92% of the potential efficiency is actually achieved at the aggregate level – the cost of the loss in efficiency is thus rather small in this context.

Lastly, in order to assess whether these results are related in any way to repetition-based learning, the bottom part of the table presents the data in terms of round-by-round behaviour. In each column, the second row aggregates all the bids posted by all the subjects in the round, the third row shows the revealed aggregate demand and the fourth row displays the revelation ratio. Each of these subjects has different induced values because the full demand curve is induced in each round, making the induced aggregate demand the same in each column (AD = 1084). The table reveals a small effect of learning on bidding behaviour, which occurs only at the very beginning of the experiment. The revelation ratio rises from 83% to 96% between the first two rounds and remains stable after that. This result illustrates the importance of practice questions in the beginning of the experiment. The result also shows that learning is not a major issue in this setting.

To sum up, the induced-value context allows us to test both theoretical properties of second-price auctions: efficiency and preference revelation. The mechanism performs worse on the first dimension than on the second, although the cost of the loss remains small. The revelation property, by contrast, is very accurately replicated by empirical behaviour. For this reason, second-price auctions have been extensively used as a mechanism to study preference revelation in the laboratory. It is worth noting that this property occurs even though most subjects were very likely to behave in the same way as those of you who had never heard about auction theory before. All these subjects might have been using rules of thumb to make up their mind, and they might also have misinterpreted the instructions. None of these behavioural elements are related to the behaviour described by theory. The main lesson is thus that the rationality assumptions that lie behind the equilibrium predictions accurately describe the outcome resulting from the environment, even though actual rationality might well not be what is behind such an observation. This (likely) discrepancy between the theoretical representation of behaviour and its actual driving forces is irrelevant in the context of this particular institution, because the institutions drive people's choices in such a way that they behave *as if* theory were descriptively right.

2.4 *Case Study*: Experimentally Designed Devices to Reduce Hypothetical Bias

Hypothetical bias in stated-preference work challenges the credibility of stated-preference methods as a tool for measuring economic values in a credible way. As Illustration 2.1 shows, hypothetical bias is best studied in the context of the elicitation of real-world goods rather than the artificial setting of induced preferences – otherwise, there is little to no difference in elicited preferences according to the monetary consequences of respondents' answers. It thus seems important to consider an elicitation context that is closer to the real-world situation of a cost–benefit analysis. This enhances what one is able to learn about the real-world behaviour thanks to the experiment – a question that will be discussed in length in Chapter 8.

One of the most common stated-preference methods used in economics is contingent valuation. This methodology uses surveys that request respondents to make decisions regarding a non-market good. Experimental researchers have tried to adapt the survey design to undermine the risk of hypothetical bias. A first possibility is to adjust, or calibrate, the answers to the valuation tasks *ex post*. An alternative is to frame the context of the individual choice to correct the hypothetical bias *ex ante*. This section will describe each of these two methods in turn. Among the *ex ante* methods, cheap talk scripts have garnered substantial attention. As such, they will be the object of a special subsection. The aim of all this research is to improve the design of preference elicitation surveys, in such a way that responses deliver a more reliable measure of true preferences in the population. They thus exemplify how experiments help public policy decision-making, which will be more systematically discussed in Chapter 9.

Illustration 2.1
Second-price auctions as a preference revelation mechanism: home-grown and induced values

An important question about the hypothetical bias phenomenon, described in Focus 2.1, is whether it actually arises as a consequence either of the mechanism itself, or of the preference revelation exercise. To study this question, Jacquemet et al. (2011) implement two sets of second-price auctions. The first one is the induced-values design described in Section 2.1. The second set of second-price auctions uses a real good: a donation to the WWF, by adopting a dolphin. In this kind of context, subjects enter the laboratory with their unknown private preferences for the good, which remains unobserved to the experimenter – hence called a *home-grown* value good. It also means that, in contrast with an induced-value design, subjects need to elicit their own preferences before answering the question. For both kinds of good, the rules of the auction remain as similar as possible – to ease comparison – but the auction is performed under two different sets of rules. In the first, subjects' earnings from the experiment are directly affected by the outcome of the auction. This condition is called the REAL treatment. In the second condition subjects are asked to behave *as if* they were directly affected by the outcome of the auction, but without any monetary consequence – hence labelled HYPOTHETICAL. The difference in revealed preferences in REAL as compared to HYPOTHETICAL measures the hypothetical bias. The results are twofold. First, there is no evidence of hypothetical bias in the induced-value context: the bids in HYPOTHETICAL are very similar to those displayed in Table 2.1. Second, the difference in revealed demand between the HYPOTHETICAL and REAL in the home-grown auction is huge. These results are in line with existing evidence reported by, e.g., Taylor et al. (2001); Vossler and McKee (2006); Murphy et al. (2010) in various experimental designs. Such results suggest that hypothetical bias is more a matter of preference formation (how subjects elicit their own preferences) rather than of preference revelation (whether self-reported preferences match the true ones). The challenge is thus to find survey designs that lead subjects to think about their true underlying preferences in the hypothetical context as seriously and deeply as they would if there were actual monetary consequences.

2.4.1 *Ex post* Methods

A famous *ex post* technique consists in calibrating down hypothetical responses – in such a way that the post-calibration values of hypothetical answers match the answers one would elicit with actual monetary incentives. Of course, the main question here is the amount of the scaling of the hypothetical responses. Many surveys in the stated-preferences literature have attempted to calculate the size of the hypothetical bias for calibration purposes. The general conclusion is that there is no golden rule for calibration. Diamond and Hausman (1994) predicted that proper calibration stipulates dividing hypothetical estimates by anywhere from 1.5 to 10. List and Gallet (2001) ran a meta-analysis on 29 studies from the literature. They found that on average subjects overstate their preferences by a factor of about three in hypothetical settings. Moreover, the amount of over-revelation appears to be good-specific and context-specific (also see Fox et al., 1998). List and Gallet (2001) found that the hypothetical bias is less important for private goods (as compared to public goods) and for willingnesses to pay (as compared

to willingnesses to accept). A possible lower bound for calibration is about 1.3, which is very close to the Diamond and Hausman (1994) lower bound.

A similar attempt at *ex post* adjustment is the use of follow-up certainty questions (Champ et al., 1997). This procedure adds a question to the survey, where respondents are asked their level of confidence in the truthfulness or accuracy of their answer to the preference elicitation survey. A threshold is then chosen, and only preferences revealed with a high enough degree of certainty, or confidence, are actually accounted for in the analysis.

2.4.2 *Ex ante* Methods

The *ex ante* methods try to build on the reasons why hypothetical bias appears to change revelation before it occurs.

A first possible reason for the poor revelation performance of hypothetical questions is that, because they are based on a hypothetical scenario, subjects do not take the valuation exercise seriously enough. Consequential procedures aim to address this issue. The procedure consists of improving the realism of the scenario (Carson et al., 2000; Cummings and Taylor, 1998). The improvement in the elicitation procedure is usually made by giving subjects the probability that their own choice in the experiment will become real. The frame underlines the fact that the participant's choice might actually impact the policy. Earlier experiments provide contrasting results on consequential procedures. Cummings and Taylor (1998) show that probabilities have to be high (greater than 0.75) to produce an effect, while Carson et al. (2002) find a coincidence with preferences elicited in the real context starting at a probability level of 0.2, which is still substantially higher than the probability level any reasonable person would assume.

In any case, even if subjects take the exercise seriously, they can still lack experience with the elicitation mechanism, or with the good to be valued. This lack of experience might lead to misconceptions, even in the case of a truthful answer to the hypothetical question. It has led some researchers to teach the valuation exercise to subjects *ex ante*, either by training them in the use of the mechanism or by increasing their knowledge of the good. In an attempt to address this last issue, Carlsson and Martinsson (2006) elicit the willingness to pay (WTP) to avoid power outages in Sweden. The WTP was expressed in an open-ended survey before and after the subjects experienced the negative consequences of the power outages. In this particular case, the WTP referred to a protection good, i.e. the right to access power without outages in the event of a hurricane. Carlsson and Martinsson (2006) obtained a somewhat paradoxical result. First, informed subjects tend to propose a 0 WTP more often than non-informed subjects. On the other hand, the answers in the subset of positive offers remain unaffected by the experience of a power outage. This paradoxical result could be explained by the fact that subjects who experienced a power outage became aware of their right to get power for free when such an event occurs. As a consequence, having faced the event of a power outage might have provided information that changes respondents' private valuation.

Focus 2.2

Preference elicitation: auctions, referenda and BDM mechanisms

While the revelation properties of the second-price auction are very attractive both empirically and theoretically, a recurrent critic against its actual use in contingent valuation surveys is its complexity. In the seminal report commissioned by the NOAA (National Oceanic and Atmospheric Administration), the panel suggests using a binary voting referendum which respondents might find more familiar and realistic. In a referendum, subjects are asked to vote for or against the funding of a public good. If a majority votes in favour, then everybody will contribute, and the public good will be provided. On the contrary, if only a minority votes in favour of the public good, then nobody will pay for the good and it will not be provided. A second attractive feature of the referendum voting procedure is that it is strategy-proof (there is no way to manipulate the outcome by distorting one's own preferences). Another popular preference elicitation tool is the Becker–DeGroot–Marschak (Becker and Brownson, 1964, BDM) mechanism. In its more standard version, a subject is asked to post a bid to buy the good. The bid is then compared to a price determined by a random-number generator. If the bid is lower than the price, the subject pays nothing and receives nothing. If the bid is greater than the price, the subject pays the randomly drawn price and receives the good. Because of this property, the equilibrium bidding strategy in the BDM mechanism is similar to what happens in a Vickrey auction: the bid affects the likelihood to buy the good but leaves unchanged the price actually paid in that case. The mechanism is thus incentive-compatible, and perfectly revealing. Noussair et al. (2004) compare the BDM mechanism and the Vickrey auction to reveal willingness-to-pay information for individual customers. For standard private goods, their results show the Vickrey auction outperforms the BDM mechanism, with fewer biases, lower dispersion of bids and faster convergence to truthful revelation.

Regarding subjects' attitudes towards the mechanism, Bjornstad et al. (1997) show that experience with contingent valuation procedure eliminates the bias. Here, experience was gained through a sequence of referenda in which participants had to vote on a proposal stating a WTP for a non-market good (see Focus 2.2 for a description of the most often used elicitation mechanisms, referenda in particular). If more than 50% of the participants voted for a given proposition, then the proposal was accepted. The good was provided and all the participants were supposed to pay the WTP. Bjornstad et al. (1997) show that a learning phase on the mechanism using real incentives strongly reduces the hypothetical bias. List (2001) studies the impact of experience by comparing experienced and non-experienced subjects. His study compared the preference elicited in a second-price auction depending on whether the card dealers were professional or not. The subjects familiar with both the good and the mechanism revealed preferences that were significantly different from those of the other subjects. The demand for professional dealers was higher than that for non-professionals: when positive, their bids were higher and their number of zero bids lower. However, experience did not succeed in overcoming the discrepancy induced by the change in the incentives context, and the hypothetical bias remained present.

Another possible problem with the hypothetical context is that subjects may face a dissonance between two competing wills. On the one hand, participants want to provide their true preferences. On the other, they want to show their support for the provision of the good to be valued. In a hypothetical context, the sending of such a message is a cost-free procedure. The dissonance minimisation (DM) procedure, introduced by Blamey et al. (1999), consists in separating the revelation of preference from the provision of support messages. The DM procedure is based on an additional response category in the survey in which subjects are explicitly asked to express their attitude towards the good. More specifically, these additional response categories clearly dissociate the respondents' support for the programme and their willingness to pay for it. For example, Blamey et al. (1999) provided respondents with the following extra response categories: 'I support the [programme] . . . but it's not worth $50 to me', 'I support the [programme] . . . but I cannot afford $50', and 'I support the [programme] but not if it requires a [fee] of any amount'. The initial study by Blamey et al. (1999) showed that DM questions elicit steeper demand functions, but they do not contrast their result obtained in a real setting.

2.4.3 Cheap-Talk Scripts

A last strand of *ex ante* methods tries to warn subjects about the hypothetical issue. In one of the first manifestations of this procedure, Bohm (1972) warns subjects involved in a public good game to avoid strategic behaviour. In a seminal contribution to the more specific field of preference valuation, the National Oceanic and Atmospheric Administration (NOAA) recommended reminding subjects of their actual budget constraint (Arrow et al., 1993). Loomis et al. (1994) tested the effectiveness of reminding subjects of their budget constraints and substitute goods, prior to elicitation. In a mail survey asking people to value old-growth forests in Oregon, they found that the effect of a reminder of this type was insignificant. Neill et al. (1994) found a similar result: the fact of reminding subjects of the value of alternative environmental goods did not change the response rates; and a similar result was found by Loomis et al. (1994). However, replications of this experiment by Kotchen and Reiling (1999) and Whitehead and Blomquist (1995, 1999) showed that this led to narrower intervals of estimated preferences when it was applied to goods with which subjects were less familiar. This approach has been systematised through the use of 'cheap-talk scripts'. A cheap-talk script provides 'persuasive' information within a social context to realign a person's behavioural expectations through communication. These scripts set the social context by explicitly revealing that people tend to overbid in hypothetical surveys (Cummings et al., 1995).

Ajzen et al. (2004) showed that the introduction of cheap-talk scripts before the decision modified the disposition of the subjects by realigning beliefs, attitudes and intentions with those in the real context. Moreover, the answers collected after a cheap-talk script are good predictors of real behaviour. However, while cheap-talk design is effective under some conditions, it is not a panacea for hypothetical bias. Evidence from the literature suggests that the length of the script is of considerable importance. For example, Aadland and Caplan (2006) found that if the cheap-talk script is short,

Illustration 2.2
An experimental comparison of correction methods

Morrison and Brown (2009) provide an experimental test of the effectiveness of the three known methods to reduce hypothetical bias: certainty scales, cheap talk and dissonance minimisation. In this experiment, students participate in a referendum where they have to decide whether the group should give a certain amount of money to the Red Cross Breakfast Club, described as an initiative to provide meals to children. Each student is given 20 Australian dollars (A$) for participation, and the amount of money to be sent to the Red Cross varies across sessions. Participants are told that if a majority votes yes, all of them will have to give the proposed amount to the Red Cross (including those having voted no). Four treatments are implemented. The first treatment is the only one with ACTUAL PAYMENTS. The three others are hypothetical so that in those treatments, students know they would keep their A$20. These three conditions are the main treatments of interest, in which preference elicitation is coupled, respectively, with CERTAINTY SCALES, CHEAP TALK and DISSONANCE MINIMISATION methods. Answers given in CERTAINTY-SCALE before the certainty question are used as the results for hypothetical estimate without correction. Certainty questions are implemented as a ten-point scale from 'very uncertain' to 'very certain'. The CHEAP-TALK treatment uses a modified version of Cummings and Taylor (1999). Finally, DISSONANCE MINIMISATION introduces four more answers, allowing students to express their support for the goal of the Red Cross Breakfast Club even if they vote against the contribution. The main results are displayed in the table below (from Table 3 in Morrison and Brown, 2009, p. 315).

Bid level (A$)	Treatment				
	ACTUAL PAYMENT	HYPOTHETICAL	CERTAINTY SCALES (limit = 7)	CHEAP TALK	DISSONANCE MINIMISATION
10	49	74	49	39	45
15	46	57	43	36	41
20	44	53	40	27	43
All	46	61	44	35	43

The experiment provides evidence of hypothetical bias: as expected, the percentage of students voting yes is greater in the HYPOTHETICAL (61%) as compared to the referendum with actual payments (46%). Dissonance minimisation appears rather efficient at correcting hypothetical bias. If calibrated at 7, the certainty scale also gives results close to ACTUAL PAYMENT. Those two methods give results which are on average 3 and 2 points away from the actual payment situation. On the other hand, the cheap-talk method produces an underestimate of the actual willingness to give. All results are more than 10 points below the actual payment. Interestingly, when asked for feedback, students report that the cheap-talk script reads as an inappropriate persuasion to vote no: 'If you're not saying no you're not being honest'.

it can actually make the hypothetical bias worse. Here, accumulated evidence favours the conclusion that a short cheap-talk script does not help to eliminate the hypothetical bias (Cummings et al., 1995; Poe et al., 2002). On the contrary, long and informative cheap-talk scripts have proven to be more valuable (Cummings and Taylor, 1999).

This success does not come without its limitations, however. In the above-mentioned experiment in which people were asked to state their willingness to pay for sports cards, List (2001) found that cheap talk did not effectively decrease the hypothetical bias for professional dealers (i.e. for agents who were well informed about the good being valued). Similarly, Lusk (2003) found that a cheap-talk script is effective in attenuating hypothetical bias only for certain classes of subject – those with less market experience or less knowledge of the good being valued. This suggests cheap talk can work as a learning booster, if the researcher provides subjects with information that under normal circumstances could only be acquired through a costly trial-and-error process. In addition, Brown et al. (2003) and Murphy et al. (2005) found that cheap-talk scripts that are long and directional work only for higher levels of provision. Carlsson and Martinsson (2006), by contrast, observe that the only effect of cheap talk is to decrease the number of zero offers, while leaving the mean value among positive offers unchanged.[7] Aadland et al. (2007) suggest that cheap talk is nothing more than an informative signal, which interacts with the anchoring effect produced by the threshold provided in dichotomous choice formats. Interestingly, this interaction results in making cheap talk drive down preferences in favour of low values but drive up preferences against high values. Based on accumulated evidence, cheap-talk scripts have to be long and detailed enough to shave the preferences towards truth revelation elicited in a hypothetical context.

Summary

The main lessons from this chapter are twofold. On the one hand, it aims to make it easier to think about the design of the experiment from the point of view of a participant – how the rules and procedures will be understood by the subjects. To that end, the chapter has shown, step by step, how an experiment proceeds from the point of view of people coming to a laboratory. This particular experiment described a second-price auction, a preference revelation mechanism with very attractive theoretical properties. First, it is perfectly revealing in theory, because the optimal strategy is to bid one's own private value for the good. Second, it succeeds in achieving the efficient allocation. On the other hand, the chapter also addresses the issue of the empirical relevance of these properties, tested through the behaviour observed in an experimental setting. The results show that the revelation properties of the mechanism are of generally good quality, while the quality of the efficiency property is more mixed. Lastly, the chapter concludes with an important policy application of the results drawn from the literature: the elicitation of preferences for non-market goods, and the design of preference elicitation mechanisms that can eliminate hypothetical bias, the main problem that arises in this context.

In reviewing this material, the chapter has also illustrated the main strengths of controlled experiments: they provide observations and control over dimensions that are otherwise (in particular, based on observational data) either or both hard to observe

[7] Ami et al. (2011) show that a neutral and short cheap-talk script can even increase the number of protest responses.

and impossible to control. The next part turns to the question of why such distinguishing features make experiments a relevant empirical method in economics. First, the point of view of the added value of experiments as compared to other empirical methods is discussed in Chapter 3. Second, Chapter 4 takes a broader perspective and relates experiments to economic theory, naturally occurring economic phenomena and the relationships between the two.

Part II

Why? The Need for Experiments in Economics

3 The Need for Controlled Experiments in Empirical Economics

As shown in Part I, experiments both widen the scope of what can be observed in an empirical situation, and are 'controlled' because the decision-making environment is built on purpose according to the objectives of the research question. In designing an experiment, and having people make decisions according to rules and with pieces of information that have all been decided on purpose, experimenters decide on what econometricians call the *data-generating process* (DGP). The aim of this chapter is to describe the consequence of this very special feature of experiments.

In a nutshell, this makes experiments well suited to help address the main challenges facing empirical economics. Empirical works in economics aim to draw general lessons from the casual evidence available in the data; e.g. what does price-sensitivity of consumers tell us about the shape of their utility function? What do differences in wages across gender tell us about discrimination? And so on. Econometrics have been developed to address questions of this kind, called inference issues. The general answer, described in Section 3.1, relies on the consistency between two sets of assumptions about the data-generating process: one about the mechanisms producing what is observed, the other about the informativeness of the statistics computed from the data. This principle is operationalised through identifying assumptions, i.e. hypotheses about how data are generated that make particular statistical treatments informative about the underlying mechanisms. The usual challenge faced in empirical economics is thus to find out the set of assumptions that best fits the unknown data-generating process inherited from the real world. Experiments reverse this challenge: they allow the data-generating process to be chosen in accordance with the empirical question to be answered.

Thanks to this property, what stand as identifying assumptions in econometrics provide a guide about how best to design experiments depending on what is to be learned from the data. An insightful source for such guidelines comes from a well-known ancestor to experimental methods in economics. Although the relevance of experiments was acknowledged only in recent decades in economic thinking, as empirical research shifted increasingly towards microeconomic-oriented works (see Section 1.2.3), a number of works have studied data very similar to experimental data for some time now. This literature focused on so-called 'quasi-experiments' (Campbell, 1969), now most often referred to as 'natural experiments'. The distinguishing feature of the econometrics of natural experiments, described in Section 3.2, is to use spontaneous changes in the institutional rules as a (quasi-)experiment. Such changes are, for instance, induced

Illustration 3.1
Labour market effects of the minimum wage: a natural experiment

Whether the minimum wage is detrimental to employment (and if so, to what extent) is a long-standing question in labour economics. The theoretical context is such that, in a world of perfect competition in the labour market, workers are paid their marginal marginal productivity at work. A higher minimum wage thus crowds out from employment workers whose marginal marginal productivity is below the minimum wage. This is only one side of the story, though. The reason why most developed countries implement minimum wages is because many labour markets are far from perfect competition. If firms would rather have market power in the labour market, then the equilibrium outcome is sub-optimal: wages are lower than under perfect competition, as is employment. Whether the minimum wage efficiently restores a balance in bargaining power between workers and firms, or crowds out productive occupations from the labour market, is thus an empirical question.

The answer to this depends entirely on the particular labour market and economy under study. In one of the most influential empirical studies on that topic, Card and Krueger (1994) exploit New Jersey's 1992 decision to amplify the federal increase in minimum wage that was adopted in the US in 1990. This decision is used as a natural experiment: it induces an idiosyncratic shock in the level of the minimum wage relative to other states, the causal effect of which can be inferred from labour market outcomes in this state compared to other states.

by modifications or new implementations of public policies (see Illustration 3.1 for an example). This literature has developed by defining identifying assumptions that are appropriate to analyse the effect of changes in the decision environment. Reviewing these identification strategies in Section 3.4 will help us think about how experimental variations must be implemented depending on the research question.

Lastly, this ability to choose the data-generating process is common to all kinds of (actual) experiments. Section 3.5 reviews the many kinds of experimental methods available in economics, often compared based on how close they are to the social situation the experiment aims to replicate. This criterion is also associated with varying abilities to actually control the data-generating process, hence giving rise to different empirical properties.

3.1 The Econometric Approach to Data Analysis

A major aim of econometrics is to inform about causal relationships between variables. To make things more concrete, our running example in this chapter will be the effect of compensation schemes on performance at work. Empirical analysis focusing on this question seeks to know what is the change in performance of workers resulting from a switch in their compensation – typically, from a fixed wage to a piece rate. To that end, econometrics makes use of statistics, but it does not reduce to it. The difference between the two is not obvious to understand. Why is it that empirical analysis in economics needs a specific set of tools, called econometrics? The answer is that structural

relationships generally are not directly observable in the data. Examples abound,[1] but to quote some of the most popular ones: the likelihood of death and the time spent at hospital are strongly positively related in any population; do such data inform us about how dangerous hospitals are for health? The number of policemen in a geographic area is often positively correlated to crime rates; does it mean one should reduce police forces to contain crime? Unemployed people who receive more help from public placement agencies generally experience lower likelihood of finding a job; do placement agencies hurt the labour market potential of job-seekers?

What these examples show is that observed relationships between variables cannot generally be trusted as a measure of their structural counterpart – correlation is not causation. Empirical correlations lie about the mechanisms generating them. In each of the above examples, there do exist forces behind the observed co-variations but these forces are not quite what simple inspection of the data suggests. Similarly, it will often be the case that higher performance is observed in firms paying a piece rate rather than a fixed wage. But without any further tool, it is impossible to know if it is so because piece rates induce higher performance, or just the reverse: that firms with a piece rate compensation scheme attract higher-performance workers (a phenomenon that could well be in operation, as explained in Focus 3.3). This discussion does not mean that data are useless, but rather that one needs to apply particular methods and reasoning to them, so as to be able to understand what is behind the observed patterns. This section summarises the framework used in econometrics to answer such questions, and highlights how it differs from statistics.

3.1.1 The Two Inferential Problems of Data Analysis

Data analysis relies on observations on a subset of the population of interest, called a *sample*. In our example, the population would include all workers of a particular kind (defined by characteristics of the workplace, specifics of the task, etc.), of which the employees of particular firms observed during a given time span are a sample. To formalise, we denote $(\mathbf{y}, \mathbf{X})$ the information available in the sample, where $\mathbf{y}$ will stand for a column vector of n individual observations on the outcome variable (e.g. the performance at work of sample employees), and $\mathbf{X}$ a matrix of n individual observations (in a row) about m input variables (in a column, such as the compensation scheme or individual observable characteristics like age or gender). A formal representation of $(\mathbf{y}, \mathbf{X})$ is the following:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \text{ and } \mathbf{X} = \begin{bmatrix} x_{11} & ; x_{12} & ; \cdots & ; x_{1m} \\ x_{21} & ; x_{22} & ; \cdots & ; x_{2m} \\ \vdots & ; \vdots & ; \ddots & ; \vdots \\ x_{n1} & ; x_{n2} & ; \cdots & ; x_{nm} \end{bmatrix}$$

Descriptive statistics are tools used to summarise the information available in the sample. The sample mean, for instance, summarises the central tendency of each variable, while covariance allows measurement of the empirical joint variations between

[1] Many additional examples of such spurious correlations can be found at `www.tylervigen.com/spurious-correlations`.

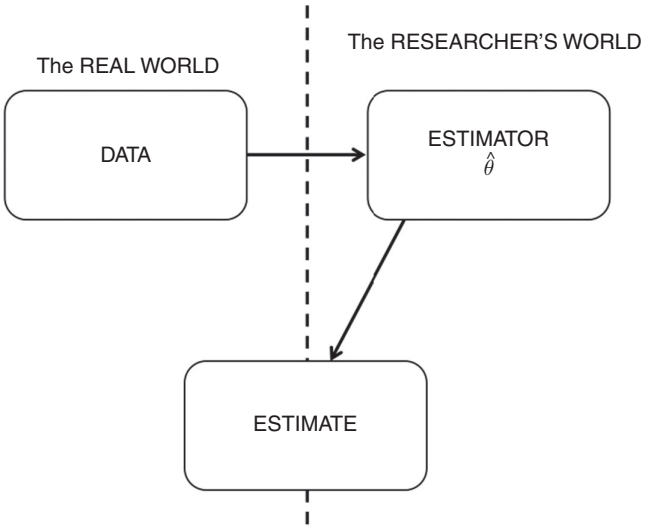


Figure 3.1 The challenge of data analysis

two variables. These are all sample quantities, which inform about the content of the variables for those individuals actually observed.

But the aim of statistics goes well beyond this objective. The main purpose is to use the information available in the sample to draw conclusions about the population characteristics. It is not the performance of those workers that actually appears in the sample that we want to understand and quantitatively characterise; but rather the behaviour of *any* worker belonging to the same population (provided the population is properly defined, something we did not do above!). This exercise is called *inference*, as the casual information available in the sample is used to infer knowledge about the population as a whole.

As an example, denote μ the mean performance at work in the target population of workers. This is something we do not observe, but we want to use sample information so as to quantify it. To that end, statistics defines *estimators*, which are procedures defined on the information available in the sample, and related to the true population parameter. For instance, the sample average of $\mathbf{y}$ defined as the procedure associated with $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ in a sample of $i = 1, \ldots, n$ observations is the estimator frequently used to inform about the mean μ. Figure 3.1 illustrates such an approach to data analysis. Data are part of the real world. Estimators, denoted $\hat{\theta}$, are defined by researchers in an effort to understand it as a whole – in such a way that the estimate can be considered as a representation of the real world as seen through the prism of the researcher's world.

An important thing to note is that such an estimator is defined for any possible sample of size n, randomly drawn from the population. From one sample to another, the particular observed values contained in $(\mathbf{y}, \mathbf{X})$ will change, because the observation units will be different as a result of the sampling mechanism. This means one can see the sample observations as n draws in the population variables. As a result, the observations contained in $(\mathbf{y}, \mathbf{X})$ are random variables, of which each particular sample giving rise to

a data set is a realisation – a particular draw in the variables' distributions. The crucial point here is that estimators are defined over sample values: as functions of random variables, they are thus random variables themselves. The application of the estimator to the actual numbers available in a given sample is called an 'estimate' or an 'estimation', and should be seen as realisations, draws, from the estimators' distribution.

These definitions allow us to state more precisely the two inference problems faced when relying on sample quantities to acquire knowledge about population quantities. The first inference problem is how the sample quantity itself is related to the population parameter of interest, i.e. what is the relationship between what we observe or compute on the data, and what we seek to measure. Imagine, for instance, that you are interested in knowing the level of income of people living in a given city. To gather this information, you stand at the entrance of a golf course nearby. You will obviously gather information that is not a relevant measure of what you are interested in. And this empirical mistake has nothing to do with the size of your sample (the number of people you will be able to meet): even if you stay long enough, your measure will never approach what you expect – because the level of income among members of a golf club is biased upwards as compared to the average income in a typical city. This first issue is a matter of **identification**: what are the relationships between the sample quantity and the true underlying parameter one seeks to measure?

But because of sample variations, what we observe will always differ from what we want to measure, notwithstanding the identification properties. It is so because, as explained above, when we compute estimations, the value of the estimator in the observed data, we work with realisations of a random variable. This question is a matter of **statistical inference**. For instance, the sample average is known to converge to the population mean if sample observations are drawn independently (in application of the law of large numbers, the probability that the sample average differs from the population mean is closer and closer to 0 as the sample size increases). The sample average will never (or barely) coincide with the population mean in any sample, whatever its (finite) size, but the higher the sample size, the closer the two will be. Using properties like this one, sample realisations can be used to characterise their population equivalent.

This two-sided inference problem is at the core of econometrics. Each side raises specific challenges. In the words of Manski (1999, p. 4), 'studies of identification seek to characterise the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations', while 'studies of statistical inference seek to characterise the generally weaker conclusions that can be drawn from a finite number of observations'.

3.1.2 How Econometrics Faces the Challenges: The Idea of Data-Generating Processes

To address these two challenges, econometrics goes even further in the distinction made in Figure 3.1 between the real world and the researcher world, as illustrated in Figure 3.2. The main novelty introduced in this figure is the inclusion of the data-generating process. There are two such data-generating processes in the figure. First,

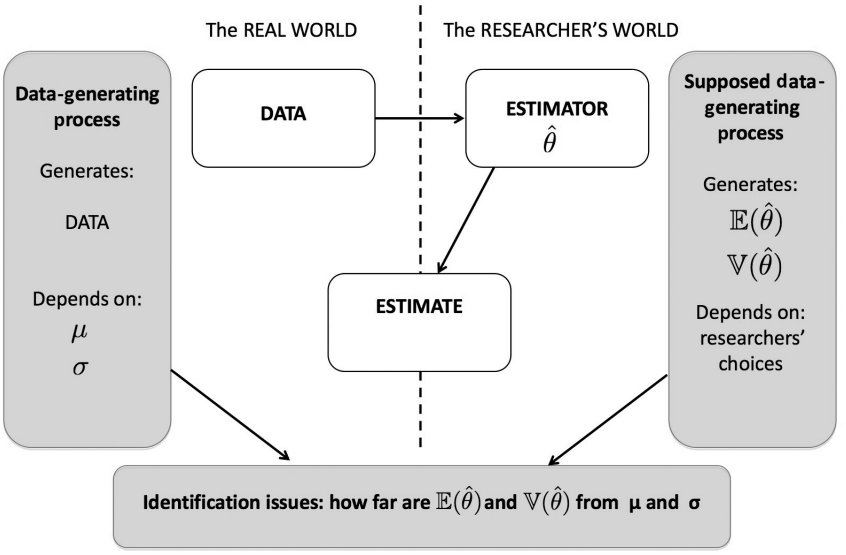


Figure 3.2 The econometric approach to data analysis

the 'true' data-generating process is the one actually producing the data observed in the real world. This includes everything that is behind the actual content of the sample: the sampling mechanism (how the decision to include a given observation from the population is made), the availability of the information (whether, for instance, variables are observed as classes, or as discrete variables, etc.), and last, all the causal mechanisms relating the variables together. By its very definition, this true data-generating process is unknown to the researcher – and impossible to observe; this is precisely the reason why empirical evidence is needed.

It is approached through the **supposed** (sometimes also labelled *assumed*) data-generating process, which gathers all assumptions made about the observed data: the functional form of the model, the assumed sampling rule, etc. It is from this supposed data-generating process that identification and statistical-inference properties of the estimator are deduced: in our running example of the estimation of the mean from a sample average, the estimator converges towards the population mean if observations are randomly drawn (i.e. the likelihood that an individual observation is included in the sample does not depend on any of his relevant characteristics). This way, the supposed data-generating process can be seen as a device to make explicit those features of the data-generating process that guarantee some particular inference properties of the estimation techniques – e.g. that observations are sampled at random, for the consistency of the sample mean.

This leads us to the main take-home lesson of this broad overview: estimators' properties are deduced from the consistency between the true data-generating process and the supposed data-generating process. Applied econometrics, from that point of view, is the art of selecting the assumed data-generating process that best fits the true one, based

on one's own understanding of what it actually is. By definition, this is not an empirical question (as any empirical analysis relies on estimators, the properties of which depend again on the consistency between their own assumed data-generating process and the true data-generating process), but rather a theoretical one: the answers will not come from the data, but from one's own understanding of the actual mechanisms at stake.

When using observational data, the true data-generating process is given, and econometric analysis of the data aims to fit for the best its main properties in order to accurately use the information available in the sample. Here lies the key difference between observational data and experiments. When designing an experiment, one actually decides on the true data-generating process: how observations are selected, how some variables are related together, what information is available and when, etc. Econometric theory thus provides the main guidelines on how to build experiments giving rise to conclusive measures. Before moving to this discussion, we illustrate the point made in this section through an application to the OLS estimator.

3.1.3 Illustration: Inference Properties of the OLS Estimator

The classical linear model linearly relates an outcome variable $\mathbf{y}$ to m covariates $\mathbf{X}$ according to:

$$y_i = \sum_i \theta_k x_{ik} + \varepsilon_i, \forall i \leftrightarrow \mathbf{y} = \mathbf{X}\theta + \boldsymbol{\varepsilon}$$

This equation is a data-generating model. From the point of view of econometrics, it literally means that $\mathbf{y}$, the outcome, is generated by a set of explanatory variables $\mathbf{X}$ combined according to the unknown parameters θ. This first component is named *the measurable* – or deterministic – part of the model, as it is made of observable variables. The way they are combined to produce outcome depends on θ, a column vector of m unknown parameters:

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix}$$

The term $\boldsymbol{\varepsilon}$ is an error, in the sense that it recovers all the variations in the actual level of $\mathbf{y}$ that are induced by mechanisms beyond the effect of $\mathbf{X}$. As a matter of fact, the equation implies that $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\theta$: the error term regroups everything that makes $\mathbf{y}$ vary and does not go through the effect of the $\mathbf{X}$s. By construction, this part of the model is *unobserved*. The error term $\boldsymbol{\varepsilon}$ is a column vector of individual error terms:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where each ε_i is defined as:

$$\varepsilon_i = y_i - \begin{bmatrix} x_{i1} & ; x_{i2} & ; \cdots & ; x_{im} \end{bmatrix} \times \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix} = y_i - \sum_{k=1}^{m} \theta_k x_{ik}$$

In a sample, one can observe the value taken by $\mathbf{y}$ and $\mathbf{X}$ for a given set of n individuals. These covariations can be used to characterise the unknown parameters θ. The most famous and widely used way of doing it is to use the ordinary least squares (OLS) estimator, $\hat{\theta}_{\text{OLS}}$. It is derived from minimising the error of the model $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\theta$ (we denote z' the transpose matrix of z):

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\theta)'(\mathbf{y} - \mathbf{X}\theta) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\theta - \theta'\mathbf{X}'\mathbf{X}\theta$$

From the first-order condition of minimising the sum of squared errors,

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \theta} = 2\mathbf{X}'\mathbf{y} - 2\mathbf{X}'\mathbf{X}\theta = 0$$

the functional form of the estimator results as:

$$\hat{\theta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

As such, the estimator is nothing more than an algebraic manipulation that maximises the fit of the model, as measured, for instance, by the R^2 – i.e. the coincidence between the observed value of $\mathbf{y}$ and the predicted value $\mathbf{X}\hat{\theta}_{\text{OLS}}$.[2]

The properties of this estimator, in terms of identification and statistical inference, result from additional assumptions on the data-generating process. To highlight them, we use the assumed relationships between the variables, $\mathbf{y} = \mathbf{X}\theta + \boldsymbol{\varepsilon}$, to write the estimator as:

$$\hat{\theta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\theta + \boldsymbol{\varepsilon}) = \theta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

According to this expression, the estimator fluctuates around the true value, θ, according to $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$. If $\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0$, $\forall i$, these fluctuations induce no systematic difference between the estimation and the true underlying parameter in such a way that $\mathbb{E}(\hat{\theta}_{\text{OLS}}) = \theta$. This defines an important identification property, called unbiasedness. Otherwise, if any correlation exists between the error term, ε, and the explanatory variables, $\mathbf{X}$, i.e. $\mathbb{E}(\varepsilon_i|\mathbf{X}) \neq 0$, then $\mathbb{E}(\hat{\theta}_{\text{OLS}}|\mathbf{X}) - \theta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) \neq 0$. This distance between the expected value of the estimator and the true parameter is called a 'bias', a systematic difference between the target and the empirical measure. Under such circumstances, the application of the OLS algebra to observed data will not deliver an evaluation of the true parameters θ combining the $\mathbf{X}s$ to generate $\mathbf{y}$. The intuitive reason for the bias is that there are some variations in $\mathbf{X}$ that are simultaneous with those in $\mathbf{y}$ (through $\boldsymbol{\varepsilon}$), not because of the causal effect of $\mathbf{X}$ on $\mathbf{y}$, however, but rather because something unobserved, hence in $\boldsymbol{\varepsilon}$, makes $\mathbf{X}$ and $\mathbf{y}$ vary at the same time. Such a phenomenon is said to be confounding: an unobserved mechanism that makes $\mathbf{X}$ and $\mathbf{y}$ vary at the same

[2] The coefficient of determination R^2 is the percentage of the response variable variation that is explained by the linear model. Formally, it is defined as $R^2 = 1 - \frac{\text{sum of squares of residuals}}{\text{total sum of squares}} = 1 - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sum_i (y_i - \bar{y})^2}$.

time is spuriously attributed to a causal effect of $\mathbf{X}$ on $\mathbf{y}$, leading to wrong conclusions and inferences.

Statistical inference, on the other hand, is deduced from the distribution properties of the n random variables ε_i in $\boldsymbol{\varepsilon}$. If these variables are identically – $\mathbb{E}(\varepsilon_i^2|\mathbf{X}) = \sigma^2 \, \forall i$ – and independently – $\mathbb{E}(\varepsilon_i \varepsilon_j|\mathbf{X}) = 0 \, \forall i \neq j$ – distributed, then it can be shown that: $\mathbb{V}(\hat{\theta}_{\text{OLS}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.[3] This quantity is the true variance of the estimator: it thus gives information on the magnitude of the variations of the realised value as compared to the expectation of the estimator, which happens to be the true value of interest if the above identification assumption is fulfilled. From one sample to another, the value taken by the estimator will vary and none of these values will coincide with the true value θ, but the range of such variations around θ is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Under the same set of assumptions, this level of precision is even the highest achievable precision among all unbiased estimators of θ (by the efficiency of the OLS): based on the available data, the OLS delivers the most informative measure of the parameter.

As this example illustrates, both identification and statistical-inference properties depend on the nature of the true data-generating process: these properties of the estimator are met only if the unobserved components gathered in $\boldsymbol{\varepsilon}$ that produce $\mathbf{y}$ actually fulfil the three conditions above. Applied to experiments, these conditions become guidelines into best practices so as to provide conclusive measures. The independence condition $\mathbb{E}(\varepsilon_i \varepsilon_j|\mathbf{X}) = 0, \; \forall i \neq j$, for instance, means that the unobserved components applying to any two observations should not be related together in any systematic manner. This condition is met in experiments if people make only one decision; but will fail (by design) if participants are asked to make several decisions one after the other. In this case, there exist several observations for which decisions are made by the same individual. All characteristics that are unobserved and specific to this person, and that belong to the decision-making process (being hungry or angry the day of the experiment, having experienced trouble getting to the laboratory, etc.), will apply to all decisions made by this person during the experiment: the unobserved components producing the outcome of several such decisions will be systematically correlated – challenging the statistical-inference properties of the statistical tools applied to these data.[4] The same reasoning applies to the identification condition, $\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0$, which is the main focus of this chapter.

3.2 Estimating Causal Effects of Treatments

Among the two issues that have to be solved in order to draw solid conclusions from observed behaviour, identification is the first-order question one must address in designing any empirical research. Statistical inference is just meaningless if identification has not been worked out, as it amounts to characterising the sample properties of the estimator as regards some unspecified parameter.

[3] For formal proofs and further discussions of the material reviewed in this section, we refer the reader to standard econometrics textbooks, such as Wooldridge (2002); Dougherty (2006).

[4] The statistical analysis of experimental data will be covered in Chapter 7.

To ease the discussion of identification issues in the context of experiments, we will rely on the so-called causal evaluation framework. This will help, in particular, to formalise the properties of the data-generating process (hence the features of experimental designs) that are crucial for achieving identification. The evaluation framework aims to measure the causal effect of a change in the decision environment on relevant outcomes. To formalise it, we frame such an empirical problem by considering two possible states of the world: $\mathcal{T} = 0$ will be the benchmark situation, while $\mathcal{T} = 1$ refers to exactly the same world except for one of its dimensions. We can think of these two states of the world as a world without a 'treatment' ($\mathcal{T} = 0$) and a world with a treatment ($\mathcal{T} = 1$), where the '*treatment*' stands for any change in the environment of which we want to measure the effect.[5] Illustration 3.2 provides an example of the kind of empirical problem this approach aims to solve. The outcome is denoted μ_{i0} in the world without a treatment and μ_{i1} in the world with the treatment. Both correspond to the true parameters of the underlying data-generating process. If the empirical question is to know the performance at work induced by different compensation schemes, μ_{i0} would be the performance under, say, a fixed wage, and μ_{i1} the performance under a piece rate. These quantities are outcomes in the sense that they are endogenous to the situation: they result from a decision or an aggregation of behaviours that is induced by an individual's reactions to the environment.

3.2.1 The Causal Effect of the Change

In this framework, the main challenge is to measure the causal effect of the treatment: the change in outcome induced by switching from world $\mathcal{T} = 0$ to $\mathcal{T} = 1$ – as discussed in Focus 3.1, this is the empirical equivalent of comparative statics in theoretical analysis. This exercise would be very easy if one could observe both μ_{i0} and μ_{i1} at the same time. It should, however, be clear from our definitions that it can hardly be the case in empirical work, as this amounts to observing the same sample unit (individual, firm, country) in both states, at the same time period, in the same sequence of events – put otherwise: both in world $\mathcal{T} = 0$ and $\mathcal{T} = 1$. As a result, if $y_i(1) = y_i(\mathcal{T} = 1)$ denotes the potential outcome individual i obtains when receiving the treatment and $y_i(0) = y_i(\mathcal{T} = 0)$ denotes the potential outcome individual i obtains when receiving the control; it is impossible to observe both $y_i(1)$ and $y_i(0)$ at the same time. A consequence is that it is impossible to measure $y_i(1) - y_i(0)$, the causal effect of the treatment, on individual i. This defines the so-called evaluation problem.

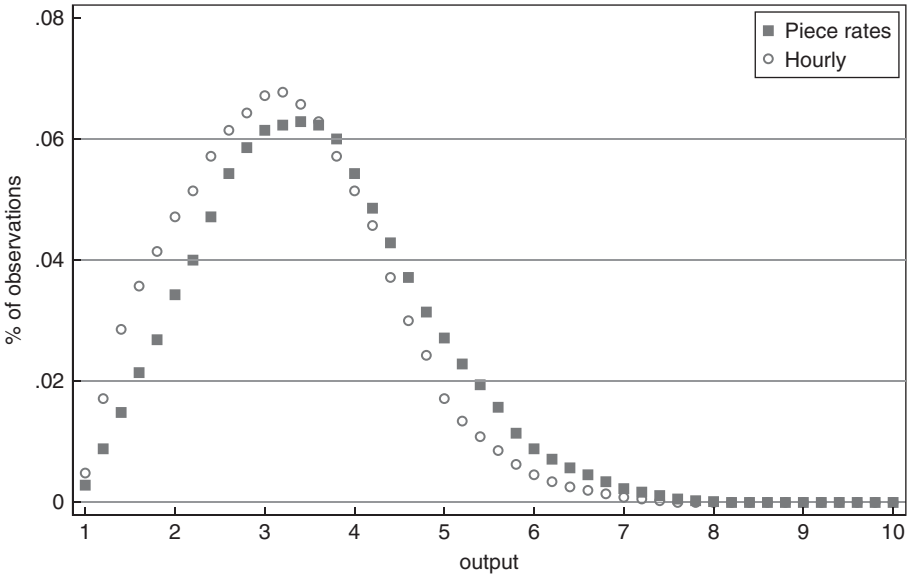
A prototypical example of this problem is the effect of education on labour market outcomes. One can observe the outcomes of different people with differing levels of education, but it is impossible to observe the outcome for one and the same individual with two different levels of education concurrently. Another example is the impact of an individual's gender on labour market outcomes, where the problem is straightforward:

[5] This formalisation is known as the Rubin (1974) causal model, from whom we borrow the title of this section. The terminology is inherited from the metaphor of experiments used in medical field: $\mathcal{T} = 1$ is a treatment tested as a cure to some illness. In that case, $\mathcal{T} = 0$ is a control, i.e. a world with no medical treatment, which refers to how patients would feel without the help of the medication in question.

Illustration 3.2
Incentives and performance: a 'natural' experiment

The aim of piece-rate schemes is to reconcile the diverging interests of the agent (or the employee) and the principal (the employer): effort at work is a cost for the agent (who cares about consumption) but benefits the principal. Piece-rate schemes achieve their goal by connecting employees' consumption to their performance at work, hence the amount of effort. This is the main reason why economic theory predicts an increased performance by switching from a fixed wage to a piece rate. Lazear (2000) relies on a natural experiment to investigate the causal impact of such a change in payment schemes on performance within a firm. The natural experiment occurred at the Safelite Glass Corporation, specialsing in automobile glass installation. Following the introduction of a new management team, the firm changed its payment scheme from hourly wages ($\mathcal{T} = 0$) to a piece rate ($\mathcal{T} = 1$). The outcome of interest is performance at work, measured by the number of glass installations carried out by an employee in a given period of time. The causal effect of interest is the change in performance induced for any given employee when work is compensated using a piece rate rather than a fixed wage.

The figure (from Figure 3, p. 1357) shows the density of performance in the firm both before and after the change in the compensation scheme. The average change in performance amounts to a rise in output of about 44% following the change. The crucial question is: does this change in performance measure the variation in output one can expect from switching an employee from one compensation scheme to the other, i.e. is this outcome a measure of a causal effect?



Focus 3.1
Causal effects in theoretical analysis and empirical works

The aim of measuring the effect of treatment variables in empirical economics is strongly grounded in economic theory. Theoretical analysis in economics consists of two complementary approaches. First, equilibrium analysis is intended to predict the state of the world that should result from a given set of circumstances, and should hence be observed in real-life situations happening under the same circumstances. *Equilibrium analysis* is based on three pillars: the existence of an equilibrium, its uniqueness and its stability. In case of multiplicity, the relevant equilibrium is the stable point around which the outcomes would converge, if all conditions remained the same. The empirical counterpart of such an approach only requires observation of the behaviour induced by a given environment replicating in the best possible way the circumstances of the model. This allows comparison of observed behaviour with the equilibrium prediction from theory – for instance: are returns to scale in the production process actually not increasing, as expected in a competitive market? The second kind of approach explicitly involves changes in the decision environment. *Comparative statics* characterise changes in the equilibrium induced by changes in the relevant circumstances. This corresponds to a systematic operationalisation of the classical *ceteris paribus* clause introduced by Marshall (1890). Comparative statics identify the relationships between variables, and the impact a change in one variable has on the outcome variable, in terms of both sign and amplitude. The econometrics of causal effects mimics the comparative-statics approach: a causal effect is nothing but a variation in behaviour induced by a change in another variable, called a treatment, in which other factors are held constant (Heckman, 2010). In measuring the change in the outcome variable before and after a change in one of the exogenous variables, the aim is to compare two different equilibrium states. This replicates the *ceteris paribus* reasoning only if the observed change is induced by the change in the exogenous variable. Anything that makes the exogenous variable change at the same time as the outcome is thus confounding, leading to biased estimation.

it is obviously impossible to observe a labour market income for one person according to whether this person is either a man or a woman – but the causal effect of gender, i.e. what are the differences in market outcomes between males and females, is a highly relevant policy question.[6]

As a result of this lack of observation opportunities, the evaluation problem consists in finding counterfactuals, i.e. empirical observations which convincingly measure what the researcher does not observe. In most instances, this is what would have happened to the observation units y_i, which we observe in the new world $\mathcal{T} = 1$, were units instead still in the benchmark world $\mathcal{T} = 0$. The accuracy of the counterfactual depends on two critical dimensions. The first is the ways in which the counterfactual observations

[6] Similarly, in the context of experiments in medical science, once the medication has been administrated to the patient ($\mathcal{T} = 1$), the researcher cannot observe what would have happened to the same patient with $\mathcal{T} = 0$. Therefore the researcher can never be sure if any change in the patient's condition stems from the medication, or from another circumstance that may have changed at the same time as the prescription. This applies to the labour market in just the same way as it does to medical research.

Table 3.1 Individuals, treatments and observations

i	$\mathcal{T}_i$	$y_i(0)$	$y_i(1)$
1	1	–	10
2	0	2	–
3	1	–	3
...	...	...	...
n	0	5	–

resemble the observations of interest. The second dimension is, of course, the kind of causal effect the researcher is seeking to identify. Before discussing each dimension in turn, we more formally specify the identification issue raised by observational data.

3.2.2 The Content of Observational Data

Observational data from natural experiments typically deliver cross-sectional information on two kinds of individuals: people who behaved in an environment where the treatment was absent (state $\mathcal{T} = 0$), and people who behaved in an environment where the treatment has been implemented (state $\mathcal{T} = 1$). In both cases, it is possible to identify those individuals who 'received the treatment' (individuals i for whom $\mathcal{T}_i = 1$) and those who didn't (individuals i for whom $\mathcal{T}_i = 0$). The observed outcome $\mathbf{y}$ in the whole sample results from the combination of the implementation of the treatment (state $\mathcal{T} = 0$ or $\mathcal{T} = 1$) and the status of the individual (treated or not). Table 3.1 shows an example of observational data for one outcome, n individuals and one treatment. Due to the outcomes delivered by the sample, the vectors of potential outcomes $\mathbf{y}(0)$ and $\mathbf{y}(1)$ are incomplete. Individuals with $\mathcal{T}_i = 1$ are missing in the former case; individuals with $\mathcal{T}_i = 0$ are missing in the latter case.

If one could observe all individuals i in a world without treatment, the model would be:

$$\mathbf{y}(0) = \boldsymbol{\mu}_0 + \boldsymbol{\varepsilon}(0)$$

where the first element, $\boldsymbol{\mu}_0$, is the vector of true parameters – or outputs – specific to this state of the world, and the second element, $\boldsymbol{\varepsilon}(0)$, is, as usual, the unobserved heterogeneity of the observational units. Similarly, if one could observe all individuals in the state of the world with $\mathcal{T} = 1$, the outcome would result from the model:

$$\mathbf{y}(1) = \boldsymbol{\mu}_1 + \boldsymbol{\varepsilon}(1)$$

Again $\boldsymbol{\mu}_1$ is the actual true parameter, while $\boldsymbol{\varepsilon}(1)$ is the measurement error due to heterogeneity. These two equations stand for the data-generating process of the outcomes in each state of the world. As mentioned above, the disturbances $\boldsymbol{\varepsilon}(0)$ or $\boldsymbol{\varepsilon}(1)$ stand by definition for any unmeasured aspect of each outcome (the whole model could be rewritten by adding explanatory variables $\mathbf{X}$, without changing the main point of the discussion: $\boldsymbol{\varepsilon}(0), \boldsymbol{\varepsilon}(1)$ will stand for any component that is influential on the outcomes, but is not explicitly measured through the $\mathbf{X}$s).

The important lesson from the discussion in Section 3.1 is that it's not the mere existence of $\boldsymbol{\varepsilon}(0)$ and $\boldsymbol{\varepsilon}(1)$ that challenges identification: there is always noise in the observed relationships between the outcome and the parameter of interest. Noise is not an issue per se, it is a necessary feature of empirical economics. What might lead to identification issues is some specific configurations of the noise, namely if the errors $\boldsymbol{\varepsilon}(0)$ and $\boldsymbol{\varepsilon}(1)$ are correlated with the treatment $\mathcal{T}$. Which kind of correlation is actually confounding entirely depends on what one seeks to measure: identification requires one not only to characterise the properties of an estimator, but first of all to define what this estimator aims to measure, i.e. the causal parameter of interest.

3.2.3 Treatment Effects Parameters

The most natural way of defining a causal effect in the above setting is as the average change in the outcome induced by the treatment for any individual from the population. This is called the average treatment effect in the policy evaluation literature. The most widely used alternative is the average treatment on the treated.[7]

The Average Treatment Effect (ATE)

This parameter measures the impact of the treatment on any individual from the population endowed with individual characteristics $\mathbf{X}$:

$$\Delta^{\text{ATE}} = \mathbb{E}\left(y_i(1) - y_i(0)|\mathbf{X}\right)$$

Intuitively, it provides a measure of the effect of moving a randomly drawn individual from no treatment to treatment, regardless of whether the individual was treated ($\mathcal{T}_i = 1$) or not ($\mathcal{T}_i = 0$). In the example of the employment effect of the minimum wage (Illustration 3.1), the ATE would be defined as the variation in the probability of employment for any individual in the population that results from a change in the minimum wage. Similarly, the ATE of a change in compensation on performance would be defined for any worker on the market.

The Average Treatment on the Treated (ATT)

The target parameter is restricted to the sub-population who receive the treatment:

$$\Delta^{\text{ATT}} = \mathbb{E}\left[y_i(1) - y_i(0)|\mathbf{X}, \mathcal{T}_i = 1\right]$$

The only difference from the previous parameter is its conditioning on the group of treated individuals.[8] The ATT measures the change in outcome for those individuals who are involved in the change. In the minimum-wage example, the ATT would now focus on the employment effects of minimum wage not for all workers, but rather for those workers who actually face a change in wage when the minimum wage changes –

[7] Heckman (2010) describes other parameters of interest like the voting criterion or more general welfare criteria. The presentation in this section closely follows the seminal survey of Heckman et al. (1999).

[8] For the sake of notational simplicity $\mathbf{y}(1|\mathbf{X}, \mathcal{T} = 1)$ and $\mathbf{y}(0|\mathbf{X}, \mathcal{T} = 1)$ will be often denoted $\mathbf{y}(1)|\mathbf{X}, \mathcal{T} = 1$ and $\mathbf{y}(0)|\mathbf{X}, \mathcal{T} = 1$. The same applies with $\mathcal{T} = 0$.

i.e. workers whose wage is just between the level of the minimum wage before and after the rise. In the compensation example, the ATT measures the change in performance induced by using a piece-rate rather than a fixed wage for those individuals who are paid a piece rate.

These two definitions make clear that these two parameters might be the same or not depending on the precise mechanism behind the change in outcome. If the individual response to the treatment is homogeneous in the population – any individual can expect the same change in outcome on average from benefiting from the treatment – then the two parameters will be the same. They are different, however, if the population is heterogeneous in terms of the response to the treatment, and assignation to the treatment is related to such heterogeneity – inducing a systematic difference in the expected effect of the treatment between individuals from the two groups.

Such a difference is driven by the existence of a relationship between the benefit from the treatment (its expected effect on the outcome) and participation in the treatment. Two kinds of implementation typically give rise to this kind of mechanism. It will be the case, for instance, if participation in the treatment is free, and those individuals who expect to benefit the most from it actually decide to get it. In the example of a change in compensation, more productive workers are likely to experience a higher raise in wage from a piece rate – in such a way that their response to higher power incentives is likely to be different from what would be the response of less productive workers. Similarly, the ATE and ATT are different quantities if the treatment targets a particular sub-population (low-wage workers, for instance) and is purposefully designed to change their outcome, rather than to improve the situation of any individual in the population. In the minimum-wage example, for instance, the ATT is likely to differ from the ATE because the labour market is strongly segmented in terms of skills, hence of wages. Individuals who earn more than the minimum wage are likely to face very little change in their employment opportunities, because the minimum wage is non-binding for the kind of job they occupy.

The ATE and the ATT both are true parameters of the distribution of the causal change induced by the treatment under study – see Focus 3.2 for a discussion of the generalisability of these two parameters, and Illustration 3.3 for an application. When they differ, the obvious question is which one we want to know and/or which will best inform on the consequences of the treatment. Unsurprisingly, the answer depends on the research question.

The ATE measures a population parameter. It thus answers questions about the likely change in the economic outcomes if the treatment is to be generalised to the whole population, or parts of the population that do not belong to the treatment group. But if the treatment is specific to those individuals who are treated, then the ATE is not very informative. For instance, a child care programme mainly targets parents with young children, and in no way aims to change the outcomes of people whose children are adults; similarly, training programmes are often designed to improve the labour market position of the long-term unemployed. In both cases, it might well be that the effect of the treatment would differ were it applied to the general population or to the target

Focus 3.2
The programme evaluation approach and the structural approach

The policy evaluation literature focuses on identification – how to best use available data to measure causal effects – from policy changes. A growing debate in this literature challenges the nature of the causal effects identified through such a 'experimentalist view of econometrics' (Keane, 2010, p. 3). The main criticism about such an approach (advocated by, e.g., Angrist and Krueger, 1999) is that the parameters identified are specific to the observed change, population, time-period, etc., i.e. they lack generality because they identify a reduced form effect. A different approach is to try to identify the mechanisms behind the causal effect – so as to achieve greater generalisability. Such a *structural* approach to evaluation explicitly specifies the mechanisms underlying individual behaviour based on the preferences, the constraints, the interactions and the sources of heterogeneity leading to a particular individual outcome. As such, the model provides a description of 'hypothetical worlds obtained by varying hypothetically the factors determining outcomes' (Heckman, 2010, p. 360). The first attractive property of this approach is to make explicit the assumptions made about the behaviour of an individual which remain implicit in the reduced-form approach (Rosenzweig and Wolpin, 2000; Keane, 2010). Second, inferences from data are based on the causal model: structural parameters leading to the reduced-form effect are estimated based on the observed variation. Such an approach thus allows one to generalise the policy effect in alternative contexts, in which determinants of behaviour are expected to be the same. The price for this increased generalisability is that the empirical analysis makes more statistical and theoretical assumptions about foundations of behaviour. Each approach thus has its own strengths and weaknesses. It is worth noting that each amounts to different specification choices, through the definition and nature of the parameters to be estimated. In both approaches to the data, though, the identification properties of the estimation depend on the structure of the data-generating process (Blundell, 2010).

individuals. But the average treatment effect is just irrelevant; what matters for both policy decision-making and academic research on the topic is the ATT. The estimation strategy must thus be adapted to the true parameter of interest – measuring the ATE without bias is of little help when the two are different and the ATT is what one actually seeks to measure.

The observational requirements of the two parameters are actually quite different. The ATT relies only on the outcomes of the sub-sample of treated individuals: $\Delta^{\text{ATT}} = \mathbb{E}\left[y_i(1)|\mathbf{X}, \mathcal{T}_i = 1\right] - \mathbb{E}\left[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1\right]$. Beyond the output under treatment for treated individuals, which is generally easy to observe, the ATT thus requires data on the outcomes of treated individuals had the treatment not been implemented – the counterfactual world for treated individuals. In the context of the example given in Table 3.1, measuring the ATT amounts to restricting the analysis to individuals who are treated – those lines in black, for which $\mathcal{T}_i = 1$. For those individuals, $\mathbf{y}(1)|\mathbf{X}, \mathcal{T} = 1$ is available. The counterfactual problem is to find a way to measure the missing values in these lines, i.e. $\mathbf{y}(0)|\mathbf{X}, \mathcal{T} = 1$.

Illustration 3.3
The need for assumptions on the data-generating process to achieve inference (even) from experimental evidence

To illustrate the lack of generalisability of experimental evidence without assumptions about the data-generating process, Manski (1999) insightfully revisits the results from the famous Perry preschool project. This natural experiment was implemented in Michigan starting in 1962. A random sample of disadvantaged black students is provided intensive educational services, while students from another random sample are used to build a control group, with no particular service. A key outcome from this experiment is that 67% of students in the treatment group were high-school graduates at age 19, while the proportion was 49% in the control group. Denote $\mathbf{X}$ the covariates of children who participated in the experiment (disadvantaged black children), $\mu_1 (= 1$ if high-school graduate by age 19, 0 otherwise) the true outcome of a child when assigned to the treatment and μ_0 the similarly defined outcome when a child does not receives the treatment. What this experiment identifies is that: $\Pr[\mu_1 = 1|\mathbf{X}] = 0.67$ and $\Pr[\mu_0 = 1|\mathbf{X}] = 0.49$. Two kinds of questions can be asked based on these results: (i) what do we learn about the effect of the programme? (ii) What would be the effect of the same programme implemented using an alternative treatment policy? The main issue in addressing the first question is that the answer not only involves the marginal distributions delivered by the experiment, but also depends on the joint distribution of the outcomes. For the sake of the illustration, consider the following joint distribution (according to which the outcomes are strongly negatively correlated):

$$\Pr[\mu_1 = 0, \mu_0 = 0|\mathbf{X}] = .00; \quad \Pr[\mu_1 = 0, \mu_0 = 1|\mathbf{X}] = .33$$
$$\Pr[\mu_1 = 1, \mu_0 = 0|\mathbf{X}] = .51; \quad \Pr[\mu_1 = 1, \mu_0 = 1|\mathbf{X}] = .16$$

This is consistent with the experimental evidence, as $\Pr[\mu_1 = 1|\mathbf{X}] = \Pr[\mu_1 = 1, \mu_0 = 1|\mathbf{X}] + \Pr[\mu_1 = 1, \mu_0 = 0|\mathbf{X}] = 0.67$ and $\Pr[\mu_0 = 1|\mathbf{X}] = \Pr[\mu_1 = 0, \mu_0 = 1|\mathbf{X}] + \Pr[\mu_1 = 1, \mu_0 = 1|\mathbf{X}] = 0.49$, and leads to graduation rates ranging between 0.16 and 1 depending on who receives the treatment. As a result, very little can be said about this question if nothing is known or assumed on the joint distribution of outcomes. This is also needed to answer the second question, i.e. to characterise the distribution of the outcome that results from policies j providing educational services to some children and not to others: $\Pr[\mu_j|\mathbf{X}]$. In this case, the answer also depends on prior information on the treatment policy that would be implemented. Manski (1999) shows that depending on the prior information used, the range of possible values compatible with the observed outcomes from the experiment is as large as [0.16; 1]. The table below (from Table 3.1 in Manski, 1999, p. 59, the computations are nothing but straightforward and we refer the interested reader to Manski (1999, pp. 60–72) for more details) shows how prior information can be used to narrow the range, and infer more informative values about the outcome of interest.

Prior information	$\Pr[\mu_j \mid x]$
No prior information	[0.16; 1]
Ordered outcomes ($\mu_1 \geq \mu_0$)	[0.49; 0.67]
Independent outcomes	[0.33; 0.83]
9/10 of the population receives the treatment	[0.57; 0.77]

In particular, the target probability lies between the two observed outcomes only if the joint distribution of the outcomes is such that the outcomes are perfectly ordered, or if the treatment is independent of the outcomes.

The counterfactual requirement for measuring the ATE is even stronger. Denoting $p = \Pr[\mathcal{T} = 1]$ the probability of being treated,[9] the ATE is defined as:

$$\Delta^{\text{ATE}} = p\,\mathbb{E}\left[y_i(1) - y_i(0)\,|\mathbf{X}, \mathcal{T}_i = 1\right] + (1-p)\,\mathbb{E}\left[y_i(1) - y_i(0)\,|\mathbf{X}, \mathcal{T}_i = 0\right]$$

The requirement for the empirical evaluation of this parameter is more demanding than for the ATT, as non-treated individuals are now included as well. In the example given in Table 3.1, this amounts to having access not only to the missing numbers for black lines where $\mathcal{T}_i = 1$, but also to the missing numbers in grey lines with $\mathcal{T}_i = 0$. On top of the observed behaviour of non-treated individuals, one also needs to find a counterfactual for the outcome of non-treated individuals had they been treated.

3.3 Identification Based on Observational Data

The formal definition of the treatment effects of interest now allows us to state more precisely the identification assumption that observational data must fulfil depending on the target parameter.

3.3.1 The Cross-section Estimator

The cross-section estimator compares the mean outcomes of treated and untreated individuals within a single period of time. The outcomes in each group are $\mathbf{y}(1)\,|\mathbf{X}, \mathcal{T} = 1$ for those who received the treatment (corresponding to the available observations in column $y_i(1)$ in Table 3.1) and $\mathbf{y}(0)\,|\mathbf{X}, \mathcal{T} = 0$ for those who did not (corresponding to available observations in column $y_i(0)$ in Table 3.1).The cross-section estimator seeks to measure the effect of the treatment based on:

$$\widehat{\Delta}^{Cross} = \frac{1}{n}\sum_{i=1}^{n} y_i(1) - \frac{1}{n}\sum_{i=1}^{n} y_i(0)$$

This expression makes clear that the estimator uses the outcome in the control group as a counterfactual to the outcome of the treated group, i.e. the estimator measures a relevant parameter only if $\mathbf{y}(0)\,|\mathbf{X}, \mathcal{T} = 0$ is an accurate measure of what the treated individual's

[9] We omit in this presentation the possible dependency of this probability on the covariates $\Pr[\mathcal{T} = 1|\mathbf{X}]$ as it is irrelevant for our discussion. This probability is also sometimes called the propensity score – the likelihood of being treated.

outcome would have been without the treatment. Formally, what the estimator delivers depends on the data-generating process according to:

$$\begin{aligned}\mathbb{E}(\widehat{\Delta}^{Cross}|\mathbf{X}) &= \mathbb{E}[y_i(1)|\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 0] \\ &= \mathbb{E}[y_i(1)|\mathbf{X}, \mathcal{T}_i = 1] + \mathbb{E}\left[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1\right] - \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1\big] \\ &\quad - \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 0] \\ &= \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 0] \\ &\quad + \underbrace{\mathbb{E}[y_i(1)|\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1]}_{\Delta^{\text{ATT}}}\end{aligned}$$

The cross-section estimator thus contains the ATT – the average treatment effect on those individuals actually treated. But it contains more, as

$$\mathbf{B}^{\text{S}} = \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 0] = \mathbb{E}[\varepsilon_i(0)|\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\varepsilon_i(0)|\mathbf{X}, \mathcal{T}_i = 0]$$

also contributes to the value taken by the comparison. This quantity measures the difference in outcome that would occur without the treatment between the two groups of individuals. If this term is non-zero, it implies that people belonging to the control group ($\forall i$ s.t. $\mathcal{T}_i = 0$) and those in the treatment group ($\forall i$ s.t. $\mathcal{T}_i = 1$) are characterised by different mechanisms producing the outcome – the data-generating process is not the same in both sub-samples. For this reason, the control group does not provide an accurate counterfactual to the treatment group outcome. As compared to the ATT, the cross-section estimator is biased with a magnitude measured by $\boldsymbol{B}^{\text{S}}$. Since this is a measure of the difference in outcome between the two groups of individuals depending on whether they will benefit from the treatment or not, this is a called a selection effect. Any such selection induces a bias in the estimation of the ATT based on the cross-section estimator.

It is worth noting that this selection effect amounts to a violation of the identification assumption because $\mathbf{B}^{\text{S}} = \mathbb{E}[\varepsilon_i(0)|\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\varepsilon_i(0)|\mathbf{X}, \mathcal{T}_i = 0] \neq 0$ implies that unobserved heterogeneity is correlated with the treatment: its distribution is different in the sub-population of individuals who will subsequently receive the treatment, and those who do not. This is typically induced by endogenous selection in the treatment, due, for instance, to the fact that programmes are targeted on people whose need for a 'treatment' is higher – resulting in lower unobserved heterogeneity as compared to those who will not receive the treatment. Focus 3.3 shows how the very same issue can arise from spontaneous choices to be treated on the part of economic agents – because people who expect the most from being treated will opt in if offered the choice.

3.3.2 Identifying Assumptions of the Cross-section Estimator

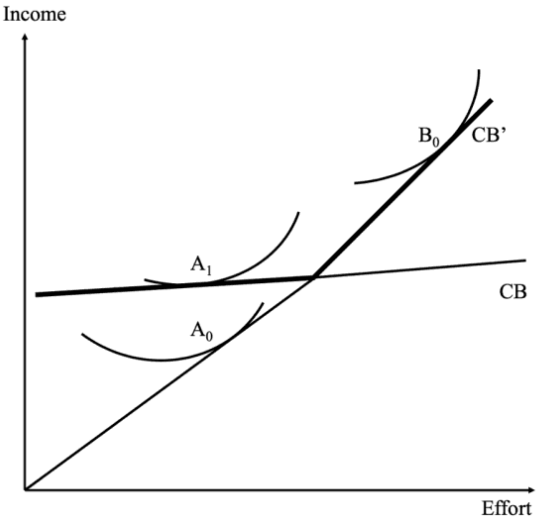
A joint product of the above result is that identification of the ATT based on cross-section comparisons will be unbiased if the data-generating process is built in such a way that there is no selection:

$$\mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 1] = \mathbb{E}[y_i(0)|\mathbf{X}, \mathcal{T}_i = 0] \Rightarrow \mathbb{E}[\widehat{\Delta}^{Cross}|\mathbf{X}] = \Delta^{\text{ATT}}$$

Focus 3.3
Incentives and performance: the confounding effect of self-selection

The gure below illustrates the choice of a worker (individual A), whose preferences are de ned over e ort (the inverse of leisure, on the x-axis) and income (y-axis). There is a trade-o between the two, as income can only be raised by increased e ort, reducing the amount of leisure. Utility thus increases towards the north-west.

The graph illustrates the e ect of switching to higher power incentives. The two straight lines are two di erent budget constraints induced by di erent compensation schemes. The rst one, CB, only weakly correlates income to e ort (reduced leisure), while the second one, CB′, is a pure piece-rate scheme o ering higher power incentives. Switching between the two moves individual A's optimal choice (the tangency between A's indi erence curves and each budget constraint) towards higher levels of e ort – what empirical analysis of the performance e ect of payment schemes seeks to quantify.



But more insights can be drawn from the gure. When moving from CB to CB′, individual A falls to a lower level of utility, as A_0 is dominated by A_1. This means A will not work under a piece-rate compensation scheme if he can choose the budget constraint – if, for instance, several rms o er the same kind of occupation but opt for di erent managerial policies, in such a way that the two budget constraints are available on the labour market. The picture is di erent for individual B, whose preferences are such that moving from CB′ to CB would lead to a lower level of utility. Individual B would thus prefer to stay on B_0, hence working in a rm o ering a piece-rate scheme. The main di erence between A and B is the shape of their preferences: in the trade-o between income and leisure, B puts more weight on income relative to e ort than A does. This may be because B experiences a lower cost of e ort, or is more e cient such that a given sacri ce of leisure leads to higher performance. The main consequence of this heterogeneity in preferences is that it is very unlikely to observe A, and very likely to observe B, in piece-rate rms. This has two consequences for empirical analysis. First, simple comparison in performance between piece-rate rms and others does not identify

any treatment effect: it mixes treatment effects and selection – the difference in performance between individuals due to their heterogeneous preferences. Second, the individual responses of A and B to higher power incentives will systematically differ, and this is related to whether or not they are actually observed in a piece-rate situation. The ATE is thus different from the ATT.

In other words, identification of the ATT by the cross-section estimator requires the affectation to the treatment to be independent of the baseline, without treatment, outcome $\mathbf{y}(0)$. As such, this condition is the identifying assumption of the ATT based on the cross-section estimator. The ATT is the parameter of interest if one seeks to measure the change in outcome experienced by people targeted by the treatment. As discussed above, this might be different from the effect obtained by giving the treatment to an individual randomly drawn from the population – or by generalising the treatment to everyone – an effect that is measured by the ATE.

The above expression is enough to show that the cross-section estimator, even under the identifying assumption of no selection on $\mathbf{y}(0)$, is a biased estimator of the ATE – just because it measures the ATT and will thus miss the ATE as soon as the two differ. The bias can be written more explicitly using the expression for the unconditional outcomes:

$$\mathbb{E}[\,y_i(1)|\,\mathbf{X}] = p\,\mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1] + (1-p)\,\mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 0]$$

so that:

$$\mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1] = \mathbb{E}[\,y_i(1)|\,\mathbf{X}] + (1-p)\,[\mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 0]$$

The same manipulations applied to

$$\mathbb{E}[\,y_i(0)|\,\mathbf{X}] = p\,\mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1] + (1-p)\,\mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]$$

lead to a similar expression for $\mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]$. The cross-section estimator is then related to the ATE according to:

$$\begin{aligned}
\mathbb{E}(\widehat{\Delta}^{Cross}|\mathbf{X}) &= \mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0] \\
&= \mathbb{E}[\,y_i(1)|\,\mathbf{X}] + (1-p)\left[\mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 0]\right] \\
&\quad - \mathbb{E}[\,y_i(0)|\,\mathbf{X}] + p[\mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]] \\
\mathbb{E}(\widehat{\Delta}^{Cross}|\mathbf{X}) &= \underbrace{\mathbb{E}[\,y_i(1)|\,\mathbf{X}] - \mathbb{E}[\,y_i(0)|\,\mathbf{X}]}_{\Delta^{\mathrm{ATE}}} + \mathbf{B}^{\mathrm{H}}
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{B}^{\mathrm{H}} = (1-p)\,&\left[\mathbb{E}\left[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1\right] - \mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 0]\right] \\
&+ p \quad [\mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]]
\end{aligned}$$

Last, simple rearrangement of the expression leads to:

$$\mathbf{B}^{\mathrm{H}} = \underbrace{\mathbb{E}[y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]}_{\mathbf{B}^{\mathrm{S}}}$$
$$- (1-p)[\mathbb{E}[y_i(1) - y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1] - \mathbb{E}[y_i(1) - y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]]$$

Even if selection on the baseline outcome is not endogenous, in such a way that $\mathbf{B}^{\mathrm{S}} = 0$, the cross-section estimator will deliver a biased measure of the average treatment effect if the second term is different from 0. The magnitude of the bias depends on the relative size of the potential treatment effects in the two groups, i.e. on the extent of the heterogeneity in the treatment effect: the higher the difference in outcomes for treated individuals as compared to the difference for those in the control group, the higher the bias (and the higher the difference between the ATT and the ATE). The only context in which the estimator identifies the ATE is when not only the baseline outcome $\mathbf{y}(0)$, but also the outcome resulting from the treatment $\mathbf{y}(1)$, are distributed independently of the affectation to the treatment – but in this case, the ATE and the ATT do not differ, because the treatment effect is homogeneous.

To illustrate selection bias and heterogeneity in the treatment effect, consider a CV-writing and job interview workshop offered at a university to foster students' success in finding a job. Students are free to choose whether or not to participate. All participants belong to the treatment group and consider that observations are also available for a random sample of non-participants from the university. A measurement strategy could be to follow both groups for a year following the programme in order to record their success in finding employment. However, the problem is that the individuals who choose to participate in the workshop in the first place may be more motivated about getting a job than those who do not. They are more likely to be hard-working students who obtain better results than the non-participants. Therefore using the treatment effect of participating in the workshop to measure how quickly students find a job may be spurious, because those students who participated would have, in any event, found a job faster than those who chose not to participate, even in absence of the workshop. The choice of participating or not is the selection effect based on individual student characteristics, making the two sub-samples difficult to compare (also see Illustration 3.4). In addition to the selection bias there is also the fact that once they have participated, the participants may benefit more from the workshop than the non-participants would have had they participated, because the more motivated participants are more knowledgeable and involved and thus draw a greater benefit from it. This represents the heterogeneity of the treatment effect.

3.4 Inference Based on Controlled Experiments

The main difference between experimental and naturally occurring data is that experiments allow us to decide on the data-generating process. First, the variation of interest is implemented on purpose in line with the research question. This is an important difference from natural experiments, for which any estimation of the treatment effect requires observations i, such that $\mathcal{T}_i = 1$, to be available. Controlled experiments, by contrast, allow us to generate any variation $\mathcal{T}$ of interest. Second, the participation decision is part of the experimental design. Identifying assumptions that have been developed in

Illustration 3.4
Incentives and performance: selection and incentive effects

The 44% increase in performance observed in the Safelite experiment (Illustration 3.2) mixes incentive effects, and selection effects associated with self-selection of workers in the firm based on their preferences after the switch to a piece-rate scheme. The table below (from Table 4 in Lazear, 2000, p. 1355) displays the differences in turnover between the hourly wage regime and the piece-rate regime (*performance pay plan, PPP*), organised according to the relative performance (decile) of workers.

	Hourly regime			PPP regime			Difference between PPP and hourly separation rates	
Decile	Separation rate	No. of obs.	St. error	Separation rate	No. of obs.	St. error	Difference	St. error
Lowest								
0	0.041	1,641	0.005	0.039	1,285	0.005	−0.002	0.007
1	0.043	1,465	0.005	0.038	1,491	0.005	−0.006	0.007
2	0.042	1,358	0.005	0.037	1,625	0.005	−0.005	0.007
3	0.039	1,245	0.005	0.037	1,691	0.005	−0.002	0.007
4	0.037	1,282	0.005	0.034	1,693	0.004	−0.003	0.007
5	0.038	1,279	0.005	0.04	1,792	0.005	0.002	0.007
6	0.025	1,223	0.004	0.03	1,777	0.004	0.005	0.006
7	0.029	1,135	0.005	0.03	1,879	0.004	0.001	0.006
8	0.03	880	0.006	0.022	2,169	0.003	−0.008	0.007
9	0.033	2,437	0.004	0.027	339	0.009	−0.007	0.009
Highest								
Overall	0.033	13,945	0.002	0.036	15,741	0.002	0.003	0.002

The simple overall effect of the change in payment regime goes from 0.033 to 0.036, but the difference is not statistically significant. The magnitude of the turnover thus remains the same. Selection, however, refers to a differential productivity of workers who leave and enter the firm due to this process. It is possible to look at this concern by focusing on those workers who work in the firm both after and before the change. On this specific sub-sample, the estimated effect is a 22% change in productivity: selection and incentive effects thus account for half the observed aggregate change in performance.

econometrics to better analyse naturally occurring phenomena thus serve as a guide for experimental practices that help identify the treatment effect of interest from the observed behaviour generated by the experiment. From the above discussion, two main devices appear to facilitate identification: one is to break correlations between unobserved components of the outcome and the change in the explanatory variable of interest; the other is to measure and thus eliminate the effect of confounding factors. Both are central to understanding how experimental methods improve the accuracy of empirical identification.

Focus 3.4
Two additional difference estimators and their identifying assumptions

The cross-section estimator uses the current behaviour of untreated individuals as a counterfactual for the treated individuals without the treatment. There exist two other very popular difference estimators, relying on alternative counterfactual assumptions.

The **before–after estimator** is based on longitudinal data – i.e. it applies when repeated observations from the same individuals are available before (at time $t = \underline{t}$) and after (at time $t = \bar{t}$) the treatment occurs. Basically, such an estimator compares the difference in the average outcome for a group of individuals before the treatment and after the treatment:

$$\widehat{\Delta}^{\mathrm{BA}} = \frac{1}{n}\sum_{i=1}^{n} y_{i,\bar{t}}(1) - \frac{1}{n}\sum_{i=1}^{n} y_{i,\underline{t}}(1)$$

This estimator thus uses the past outcome of the treated group as a counterfactual to the outcome of the group being treated at the time. Based on the data-generating process, the identification achieved by the estimator is

$$\begin{aligned}\mathbb{E}(\widehat{\Delta}^{\mathrm{BA}}|\mathbf{X}) &= \mathbb{E}[\,y_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1, t = \bar{t}] - \mathbb{E}[\,y_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1, t = \underline{t}] \\ &= \Delta^{\mathrm{ATT}} + \mathbb{E}[\,\varepsilon_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1, t = \bar{t}] - \mathbb{E}[\,\varepsilon_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1, t = \underline{t}]\end{aligned}$$

The second term is the variation in unobserved heterogeneity in the group of treated individuals, from before ($\boldsymbol{\varepsilon}(1)_{t=\underline{t}}$) to after ($\boldsymbol{\varepsilon}(1)_{t=\bar{t}}$) the treatment. By construction, any permanent (typically, individual-specific) heterogeneity is eliminated. But any difference occurring over time will not be eliminated, and biases the estimation of the treatment effect – because the treatment variable is then correlated with time-varying unobserved heterogeneity. The identifying assumption of the BA estimator is thus that there is no unobserved change over time inducing a variation in the outcome beyond the treatment itself. Otherwise, the bias comes from the fact that the estimator attributes any change over time in the outcome to the causal effect of the treatment.

The **difference-in-difference estimator** relies on both cross-sectional and longitudinal data, and estimates the treatment effect for a given group of individuals as:

$$\widehat{\Delta}^{\mathrm{DD}} = \left(\frac{1}{n}\sum_{i=1}^{n} y_{i,\bar{t}}(1) - \frac{1}{n}\sum_{i=1}^{n} y_{i,\bar{t}}(0)\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_{i,\underline{t}}(1) - \frac{1}{n}\sum_{i=1}^{n} y_{i,\underline{t}}(0)\right)$$

It amounts to using the past difference between treated and untreated individuals as a counterfactual of their current difference. The aim of this double difference is to eliminate confounding issues related to problems of both timing (as in the BA estimator) and selection (as in the cross-section estimator):

$$\mathbb{E}(\widehat{\Delta}^{\mathrm{DD}}|\mathbf{X}) = \Delta^{\mathrm{ATT}} + \mathbf{B}^{s}_{\bar{t}} - \mathbf{B}^{s}_{\underline{t}}$$

As a result, the identifying assumption of the difference-in-difference estimator is also known as a 'parallel trend assumption': the unobserved difference between individuals from the control and treatment groups should remain unchanged or, put otherwise, the change in outcome in the control group must replicate the change that would have occurred for treated individuals, without the treatment.

3.4.1 Identification through Randomisation

The above discussion of both the OLS estimator and the econometrics of treatment effects points to one crucial condition for identification: the exogeneity of the explanatory variables, i.e. that there is no correlation between the unobserved component of the outcome and the variables of interest. The most natural way to avoid correlation is randomisation. It is quite straightforward to see in the identification condition of the OLS estimator, derived in Section 3.1.3:

$$\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0,\ \forall i \Rightarrow \mathbb{E}(\hat{\theta}_{\text{OLS}}|\mathbf{X}) = \theta$$

Identification is achieved if the distribution of the error is not correlated with the values taken by the $\mathbf{X}$. In the extreme case, this would obviously be obtained if one can decide on a purely random basis the value of the explanatory variable for each and every individual. If, for any i, the value taken by $x_{i1}, x_{i2}, \ldots, x_{im}$ results from a random draw then it is mechanically the case that this value is independent of the unobserved component of any outcome variable. Think, for instance, of the measurement of gender-specific behaviour in groups (of which Illustration 3.5 provides an example). The aim of such a study is to measure the causal effect of the gender composition of a group on male and female members. The challenge faced when trying to measure such an effect on observational data is the endogeneity of the composition variable: people may decide whether or not to belong to different groups according to their gender composition based on their specific ability to accommodate such circumstances. This induces a correlation between the composition variable and some unobserved component of the outcome of interest. While gender obviously cannot be randomly chosen (even in an experiment!), the group composition faced by male and female subjects can easily be randomised. This achieves identification of the causal effect of interest.

This same principle applies as well to the estimation of treatment effects in the policy evaluation framework. As shown above, selection bias and heterogeneity of the treatment effect arise when individuals can freely choose their treatment groups, i.e. if people sort themselves into treatment groups according to their characteristics and preferences. The way to counteract this phenomenon is again randomisation, applied to the treatment participation groups. This amounts to choosing randomly whether individuals belong to the control or to the treatment group. Under this rule (often called random assignment to the treatment), the value of the treatment variable is decided by a coin toss. By construction, this implements the identifying assumption of the cross-section estimator, i.e. the average non-participant in the programme – $\mathbb{E}[\mathrm{y}_i(0)|\mathbf{X}, \mathcal{T}_i = 0]$ – obtains the same non-treatment outcome as the average participant in the programme – $\mathbb{E}[\,\mathrm{y}_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1]$). Put otherwise:

$$\mathbb{E}[\,\varepsilon_i(0)|\,\mathbf{X}, \mathcal{T}_i = 1] = \mathbb{E}[\,\varepsilon_i(0)|\,\mathbf{X}, \mathcal{T}_i = 0]$$

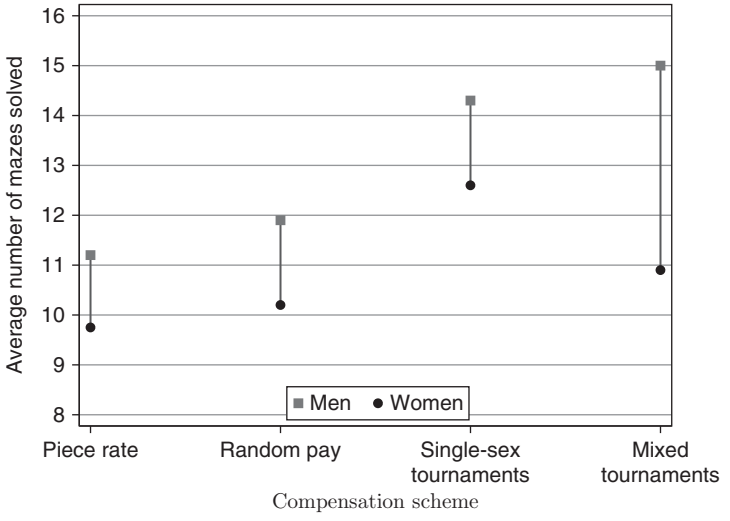
A simple comparison in outcomes between treated and untreated individuals (based on the cross-section estimator) is thus enough to measure the ATT (see Illustration 3.6 for a simple application in the field). But randomisation in fact does more: because participation in the treatment is not at all based on unobserved heterogeneity, it is also the case that

Illustration 3.5

Gender differences in competitiveness: experimental evidence from exogenously chosen composition of groups

Gender differences in labour market outcomes are well documented facts in most countries. In order to better understand the behavioural reasons behind such differences, Gneezy et al. (2003) experimentally investigate whether competition induces gender-specific behaviour. More specifically, the experiment considers tournaments in which the gender composition of the team is either all females or males, or mixed with exactly half males and half females. The task in the experiment is to solve mazes. The baseline provides a control on possible performance differences between gender, by having the task paid according to a piece rate. A fourth treatment aims to provide a control on the sole effect of the uncertainty about compensation faced in tournaments due to the need to anticipate other members' performance. In this treatment, compensation is individual but uncertain – one individual is randomly chosen to be paid according to the piece rate. The figure below (from Figure 3, p. 1062) displays the main results obtained in the experimental treatments.

For each treatment in abscissa, the figure displays the performance of males and females as well as the gender gap. The gender differences are insignificant in both the piece-rate and the random pay scheme. Both males and females positively react to competition, since both kinds of tournament elicit higher performance. But group composition has a strong effect on the gender gap: females do much worse when competing with males rather than females, while males do slightly better.



$$\mathbb{E}[\,\varepsilon_i(1)|\,\mathbf{X}, \mathcal{T}_i = 1] = \mathbb{E}[\,\varepsilon_i(1)|\,\mathbf{X}, \mathcal{T}_i = 0]$$

Randomisation also eliminates any heterogeneity in the treatment effect, in such a way that there is no difference between the ATT and the ATE. Random assignment allows us to measure the expected effect of the treatment on any individual randomly drawn from the population.

Illustration 3.6
Piece rate: a field experiment

Shearer (2004) reports on a field experiment aimed to measure the impact on productivity of piece-rate payment mechanisms. The experiment is implemented in the tree-planting industry in British Columbia (Canada). The industry has a number of characteristics that facilitate the study of interest: daily compensation in the tree-planting industry varies regularly according to the properties of the land area where the work occurs. This makes the sudden change in compensation (from the usual hourly wage to a piece rate) more natural for workers. Productivity is also easy to measure and can be simply computed as the number of trees planted during a given period of time. In order to run the experiment, three areas of varying levels of difficulty were selected, and each is randomly subdivided into two parts. The two parts define compensation regions: one has a fixed wage (FW) announced and paid, while the other part has a piece-rate payment (PR). Nine workers were randomly selected for 120 workings days: each of them is reallocated each day to a land area and a compensation region, in such a way that 60 daily observations are available in each compensation region. The table below (from Table 3, p. 518) provides the descriptive statistics from the experiment for each planter.

Planter	Observations	Total	PR	FW	Difference
1	16	1127.50	1275.00	980.00	295.00
2	12	1098.33	1220.00	976.67	243.33
3	12	1226.67	1430.00	1023.33	406.67
4	16	992.50	1000.00	985.00	15.00
5	12	1163.33	1266.67	1060.00	206.67
6	4	1330.00	1470.00	1190.00	280.00
7	16	1121.25	1165.00	1077.50	87.50
8	16	1157.50	1255.00	1060.00	195.00
9	16	1252.50	1420.00	1085.00	335.00

Thanks to the random allocation to groups, simple comparisons provide an estimate of the average treatment effect. It amounts to a 20% increase in productivity induced by the change in incentive.

It is important to be precise on the definition of the population – and generalise our understanding of the ATT/ATE. The above relations make it clear that identification is specific to the experimental population (from which individuals from both groups are drawn). The estimation is an ATE in the sense that the experiment measures the average expected change for any individual randomly drawn from this population. But it does not mean it is the effect of the treatment on any human being among those who do not participate in the experiment: if such people are likely to react differently to the treatment, then the experiment obviously fails to identify what the treatment effect would be in this specific population. Whether or not such a difference challenges the lessons drawn from experiments depends on the definition of the population on which inference is made. This is a matter of intensive debate in the literature, further discussed in Chapter 8. But in terms of the definitions above, this has nothing to do with a biased estimation of the effect of interest; it is rather a matter of a heterogeneous treatment effect – the experimental ATE is in fact specific to the experimental

subject pool. The effect is accurately measured in the experiment, but it might not generalise to other populations because the treatment effect is specific to the experimental population.

3.4.2 Identification through Control Variables

One distinguishing feature of experiments discussed in the previous chapter is their ability to widen the set of available measures – improving the set of research questions that can be addressed based on experimental data. This also plays an important role in the identification properties of experiments, as it widens the set of control variables that can be used to enhance identification. In econometrics, control variables are individual specific information that are accounted for to improve identification, rather than because their effect actually belongs to the research question. Despite this strong difference in nature, concretely speaking they are nothing but additional explanatory variables – typically, age, income, occupation, etc. A typical example of control variables is their use in discontinuity designs. Consider a policy that applies according to a threshold – e.g. additional benefits offered to people whose income is below a target – in a such a way that people who benefit from the policy are selected as regards to the outcome variable. The distance to the threshold is a control variable achieving identification: conditional on being in a close neighbourhood around the threshold, the treatment status can be assumed to be exogenous, i.e. people around the threshold can be assumed to fall below or above for purely random reasons (see Black, 1999, for an application to the measurement of parents' willingness to pay for the quality of schools, based on discontinuities at district borders).

Note that in all previous discussions, the explanatory variables $\mathbf{X}$ enter the identifying assumptions in two different ways. They enter directly in the conditioning, reflecting the fact that their effect cannot be identified unless they are exogenous – uncorrelated with the unobserved components of the outcome. The consequences in terms of identification are the ones discussed above: randomising participation in treatments based on individual specific characteristics is enough to achieve identification.

But individual specific variables also enter indirectly the identifying assumption, through the unobserved component itself. Remember that the noise component of the model, $\boldsymbol{\varepsilon}$, is residual in nature: it stands for the part of the outcome data-generating process that is not accounted for by the measurable part involving $\mathbf{X}$. In the OLS specification, this has been noticed in Section 3.1.3 from a simple rearrangement of the linear equation leading to $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\theta$. By construction, any component leading to $\mathbf{y}$ for which an empirical measure is available is thus eliminated from $\boldsymbol{\varepsilon}$ as soon as it is included in $\mathbf{X}$. This opens an additional strategy to improve identification. Adding measures to the model will reduce the noise, hence undermining the scope of possible confounding factors in the error term. It is worth insisting again on the fact that such control variables and additional measures do not improve identification just because they reduce the noise. Noise itself is a matter of statistical inference, not of identification. The circumstances in which they improve identification are when unmeasured dimensions that are correlated with the observables of interest become measured. For instance, the case study below (Section 3.4.4) discusses the potentially confounding

effect of risk attitudes in the evaluation of alternative compensation schemes. In a study on observational data, such an individual specific characteristic is likely to be unmeasurable and belong to the error $\boldsymbol{\varepsilon}$ of a model estimating the determinants of performance at work. The experimental context, by contrast, allows us to collect data on this dimension (using specific elicitation methods which are the topic of the case study presented in Section 7.4). Once a measure of individual risk attitudes is available, its effect on the outcome no longer belongs to the unobserved component of performance. The model involves an error $\tilde{\boldsymbol{\varepsilon}}$ identical to $\tilde{\varepsilon}$ but excluding risk attitude. If risk attitudes are actually confounding in the study, identification is thus recovered thanks to their explicit measure.

3.4.3 Enhanced Inference Thanks to Control

A last possibility to achieve identification is to rule out confounding variations in the noise, by blocking their value to a specified level. The idea is fairly simple: consider a situation with an unobserved component, z, generating endogeneity – hence belonging to the noise $\boldsymbol{\varepsilon}$. In the policy evaluation framework, this means that variations in the value of z induce variations in the value of both y and $\mathcal{T}$ in such a way that $\mathbb{E}[z \in \varepsilon_i(0) | \mathbf{X}, \mathcal{T}_i = 1] \neq \mathbb{E}[z \in \varepsilon_i(0) | \mathbf{X}, \mathcal{T}_i = 0]$, because $\boldsymbol{\varepsilon}(0)$ contains different values of z when looking at different values of $\mathcal{T}$. Randomisation solves this issue by letting z vary from one individual to another, $z_i \neq z_j,\ i \neq j$; but by breaking the correlation between unknown causes of the outcome and the target variation $\mathcal{T}$, in such a way that $\mathbb{E}[\varepsilon_i(0) | \mathbf{X}, \mathcal{T}_i = 1] = \mathbb{E}[\varepsilon_i(0) | \mathbf{X}, \mathcal{T}_i = 0]$. Identification is thus achieved thanks to statistical balance in unobserved components between groups: the values of z are different between different individuals, but randomisation makes it more and more likely as the sample size becomes bigger that the distribution of these values in subgroups defined by $\mathcal{T}$ is the same. Consider instead a design holding z constant at a given value: $z_i = \underline{z},\ \forall i = 1, \ldots, N$. Because the value of z now is the same for all individuals, there is no longer any difference in the unobserved component of y it induces between the two sub-samples; since there is no variation in z, there is no confounding variation in the data. The big difference is that the sample now is exactly balanced, i.e. $\{z | \mathbf{X}, \mathcal{T} = 1\} = \{z | \mathbf{X}, \mathcal{T} = 0\} = \underline{z}$.

This property implies that using control as an identification strategy comes with an improvement in statistical inference as compared to randomisation. Remember that statistical inference refers to the uncertainty about the value of the target parameter measured thanks to an estimate, due to variations in the sample. This is again summarised in the noise component of the econometric model: the target parameter is imperfectly measured based on the observed relationship between the outcome variable and the exogenous regressors because this observed relationship is only an imperfect signal of the true relationship going through the unknown parameters. In the context of the linear model of Section 3.1.3, $\mathbf{y} = \mathbf{X}\theta + \boldsymbol{\varepsilon}$ implies that $\hat{\theta}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\theta + \boldsymbol{\varepsilon}) = \theta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$: the more noisy the data, the larger the estimator variations around the true parameter. In this example, this results in a precision equal to $\mathbb{V}(\hat{\theta}_{\text{OLS}} | \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, which is increasing in the variance of the noise. The more unobserved dimensions are set to a unique, constant, value, the lower this quantity will be.

For the sake of the illustration, consider again two different components embedded in the noise of this linear model: an unobserved factor $\mathbf{z}$, and everything that remains beyond its effect on y, denoted $\tilde{\boldsymbol{\varepsilon}}$. Writing this noise as $\boldsymbol{\varepsilon} = \theta_z \mathbf{z} + \tilde{\boldsymbol{\varepsilon}}$ – i.e. θ_z measures the extent of the variations in y that come from variations in z. Now consider two data-generating processes. In the first, $\mathbf{z}$ freely varies from one observation to another. The noise of this empirical model is such that $\sigma_\varepsilon^2 = \theta_z^2 \mathbb{V}(z) + \sigma_{\tilde{\varepsilon}}^2$. Now consider the other extreme, in which $\mathbf{z}$ is still unobserved (typically because it is impossible to build a measure of it) but is held constant to a given value, $z_i = \underline{z},\ \forall i$. In this sample, $\mathbb{V}(z) = 0$ so that $\sigma_{\varepsilon_0}^2 = \sigma_{\tilde{\varepsilon}}^2 < \sigma_\varepsilon^2$: the estimates drawn from such a sample will be closer to the target parameter, hence more informative, because data are less noisy. As a result, the sample quantities delivered by experimental data are more informative the more unobservable dimensions are held constant between observations – which also improves identification as soon as they are possibly confounding.

3.4.4 *Case Study:* The Incentive Effects of Tournaments: Evidence from the Laboratory

The main limitation of the use of piece-rate incentive schemes is that they are not renegotiation-proof when effort is non-verifiable – i.e. when the level of performance cannot be proven based on evidence (Malcomson, 1984). In this kind of situation, the employer always has an incentive to renegotiate the work contract *ex post*, once the performance of the employee becomes known, so as to avoid paying high wages. Tournaments are an alternative compensation scheme that are robust to this issue.

The principle of a tournament (introduced by Lazear and Rosen, 1981) is to rely on relative, rather than absolute, performance. To formalise the comparison between tournaments and piece rate, assume all workers i in a firm have a production function $y_i = g(\ell_i) + \varepsilon_i$, where effort ℓ_i is exerted at cost $c_e(\ell_i)$.[10] If employees work under a piece-rate payment mechanism, denoted w, then the level of incentives is $w.y_i$ so that performance is chosen according to $c_e'(\ell_i^*) = w.g'(\ell_i^*)$. Since the wage is an increasing function of effort at work, the optimal level of effort increases in the piece rate. A tournament relies on different 'prizes', for instance A and a such that $A > a$, that are distributed according to the rank of the employee in the distribution of efforts. For instance, worker i gets A if $y_i > y_j, j \neq i$ but will get a otherwise. In the context of the simple production function above, the tournament provides incentives according to the probability of winning the biggest prize $\pi(\ell_i, \ell_j) = \Pr[\varepsilon_i - \varepsilon_j > g(\ell_j) - g(\ell_i)]$. Based on this probability, it is possible to replicate the optimal effort induced by any piece rate through an appropriate choice of the prizes offered in the tournament. The main difference between the two is that the total amount of the wage bill is determined *ex ante*: there must be a winner who gets the highest prize. The tournament is thus renegotiation-proof.

Whether or not tournaments empirically achieve the same level of incentive as piece rate, however, remains an open question. Bull et al. (1987) provide evidence aimed at comparing the incentive effects of these two incentive schemes. In the tournament treatment, subjects are told that they have been assigned a partner in the room, whose identity

[10] $g(\ell_i)$ is assumed to be increasing, continuous and concave in ℓ_i, and $c_e(\ell_i)$ is assumed to be increasing, continuous and convex in ℓ_i.

remains hidden. Subjects are then asked to pick a number between 0 and 100, called the 'decision number', standing for the effort variable. The cost function is set equal to ℓ_i^2/b, where ℓ_i is the subject's chosen number and b a parameter. Rather than asking subjects to do their own computations, subjects are provided with a 'cost list' with the individual cost of each number between 0 and 100. Last, subjects are asked to draw a random integer between $-h$ and $+h$ (0 included) from a box of bingo balls, standing for the noise between performance and effort. The amount of production is set equal to the sum of the two numbers. The experimenter computes each subject's production, and announces and rewards the subject pair with the highest number. The parameter values are chosen in such a way that the equilibrium effort (ℓ^*) in this baseline tournament is equal to 37 for each individual. In a benchmark experiment, the compensation scheme is set to a piece rate tuned in such a way that it replicates the same equilibrium level of effort.

The upper part of Figure 3.3 compares the level of effort observed in the experiment depending on the compensation scheme. On each graph, the straight line represents the

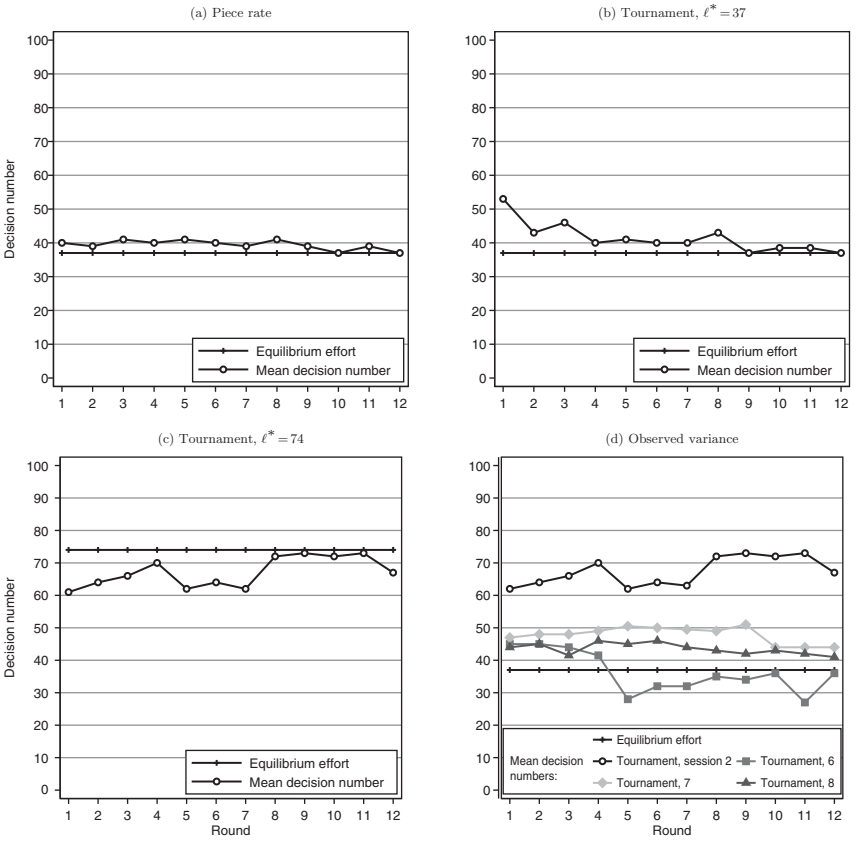


Figure 3.3 Incentive effects of tournaments

Note. Observed effort over time in the piece-rate treatment (a), the tournament treatment with equilibrium effort equal to 37 (b) and the tournament treatment with equilibrium effort equal to 74 (c). d shows the variance in effort observed over time in all treatments.

Source: Bull et al. (1987, p. 17–20, Figures 1–3, 11).

equilibrium effort and the joint dots are the mean decision numbers chosen in each round. First, the results clearly show that tournament and piece rate achieve the same convergence rate towards the equilibrium effort. In order to double check this coincidence between the theoretical prediction and empirical behaviour, a third experiment is implemented in which the tournament is designed to induce an equilibrium effort equal to 74. Observed behaviour further confirms the incentive effect of the tournament scheme: as shown in Figure 3.3.c, the level of effort chosen by the subjects matches the new equilibrium value. In both instances, the rate of convergence, however, seems lower in the tournament treatments as compared to the piece rate. Figure 3.3.d aims to investigate the reasons why it happens: rank-order tournaments exhibit a very high variance in effort between subjects, while piece-rate schemes have a rather small variance.

Such a variance in effort may reflect the need to account for other workers' behaviour: the decision in a tournament is a function both of a worker's preferences and skills and of their co-workers' behaviour – heterogenity in skills and preferences is likely to make coordination at work harder, as shown by Meidinger et al. (2003). This might affect not only the incentive properties of tournaments, as shown above, but also their selection effects. Eriksson et al. (2009) aim to study the determinants of self-selection into tournament schemes. The experiment involves two treatments, a benchmark and a choice design. The benchmark experiment allocates subjects randomly to either piece-rate payment or a tournament, while the second treatment allows subjects to choose *ex ante* how they want to be paid in each period.

Figure 3.4 focuses on the dispersion in effort observed under a tournament compensation scheme: the larger the grey areas, the further are quartiles of the effort distribution from its median. Self-selection is the only difference between the left-hand-side figure and the right-hand-side one: subjects are randomly assigned to the tournament in the benchmark, while they freely opt for it in the choice treatment. Self-selection drastically decreases the variance of effort: half the variance disappears when people self-select

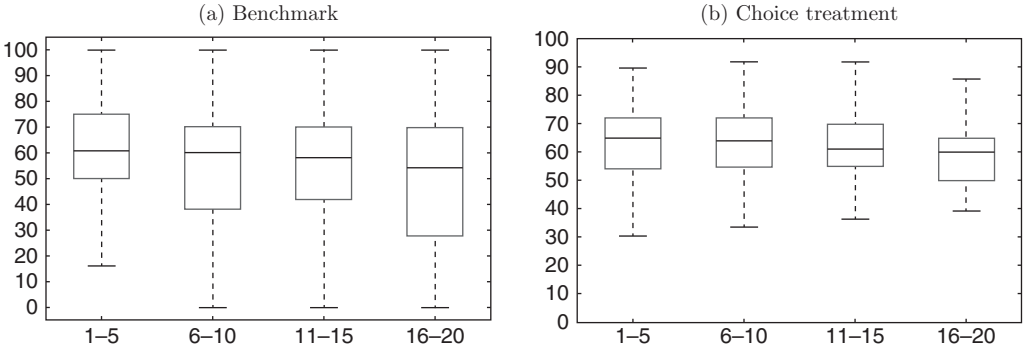


Figure 3.4 Dispersion of efforts in tournaments

Note. Observed effort in the tournament in each treatment, according to the time period in the experiment (in abscissa). The straight line shows the median effort, the box provides the quartiles and the horizontal lines show the adjacent values.
Source: Eriksson et al. (2009, p. 538, Figure 2).

themselves according to their preferences and abilities. Based on additional control variables, the authors also observe that those who select the tournament have particular preferences: it is chosen half the time, by the least risk-averse subjects. Illustration 3.5 shows that the same kind of phenomenon applies according to gender.

3.5 From the Laboratory to the Field: An Overview of Controlled Experiments in Economics

The ability to better achieve identification through control over the data-generating process is common to all (actual) experimental methods. This feature is shared, in particular, by the two types of controlled experiment that are often referred to in the literature: laboratory and field experiments.

3.5.1 The Many Forms of Field Experiment

A field experiment is commonly defined as an experiment that produces observations from (i) a random allocation of individuals to the treatment, but (ii) in a '*natural*' or '*real*' environment (field experiments are also sometimes called 'random controlled trials' in the literature). The latter underscores the main difference between a laboratory experiment, in which social phenomena are observed in highly artificial circumstances – implementing 'the sterility of the laboratory experimental environment' according to Harrison and List (2004, p. 1009) – and a field experiment, in which individual behaviour is observed in its naturally occurring environment. The field is thus characterised by the fact that subjects take decisions in the same social context as they would under normal circumstances; the information they receive makes use of this social context which the experiment reproduces; they are familiar with the rules they are dealing with, etc.; in a nutshell: all features ensure that the environment remains as 'realistic' as possible (see Illustration 3.6).

While the difference in nature between the two kinds of environment is easy to understand, there are many dimensions that make observed decisions close or far from their naturally occurring circumstances. In an attempt to summarise the relevant dimensions, Harrison and List (2004) characterise experiments according to the following six items:

1. the nature of the subject pool,
2. the nature of the information that the subjects bring to the task,
3. the nature of the commodity,
4. the nature of the task or trading rules applied,
5. the nature of the stakes,
6. the nature of the environment that the subject operates in.

Even if one considers that there is a choice between only two options for each of these dimensions, one closer to the naturally occurring environment and the other closer to the

sterility of the laboratory, this classification defines $6 * 6 = 36$ different kinds of experiment. Except for the two extreme cases, where all six dimensions are set to either their naturally occurring or their artificial value, none of the remaining 34 'types' of experiment are clearly a field or a laboratory experiment. More recently, Gerber and Green (2012, p. 11) define the 'fieldness' of an experiment based on the following criteria:

1. authenticity of treatments,
2. identity of participants,
3. the nature of the context,
4. the outcome measures.

This narrower classification still leads to at least 16 different kinds of experiment, with varying relationships between the phenomenon of interest and the naturally occurring environment. This points to the fact that 'field' or 'laboratory' is hardly a label one puts on experimental evidence to indicate the main focus of the research. In terms of the content of the experimental design, whether an experiment should be classified as a field or laboratory experiment rather defines a continuum with an increasingly close match between the social situation of interest and the experimental task.

3.5.2 A Guide to Choosing between Lab and Field Experiments

Experiments along the continuum nonetheless feature varying properties in terms of identification and the overall quality of the measures they generate. According to a well-known econ joke, if you ask three economists what they think about an economic issue, you'll get five different opinions. Having a look at experimental papers published in the academic literature, the same reaction seems to apply to the question whether laboratory experiments are superior to field experiments or vice versa. Having recognised that neither laboratory experiments nor field experiments are self-contained methodological categories, it is easier to understand what the answer is: it depends entirely on the research question, hence the social situation and the treatment effect one wants to identify. Along the continuum between field and laboratory experiments, each choice amounts to a trade-off, whose main dimensions are described below.

The first set of trade-offs involves practical aspects and the ease of implementation. Its cost is the most obvious issue. The closer an experiment is to a pure field experiment, the more expensive it is likely to get. Experiments occurring in the field are extremely costly in terms of the resources they consume. They take a lot of planning, time and money before they can be implemented. The first step is to get the agreement of the real-world institution where the experiment will take place, and interact on-site about the details of the experiment. It is also necessary to recruit local research assistants in order to find participants and conduct the experiment. The experiment may run over a very long period of time before the researcher is able to observe and gather data on its outcomes. In addition, the researcher has to ensure that the research design respects local customs and habits. All this can make the implementation of a field experiment a very lengthy and costly process.

Second, the choice between the field and the laboratory implies a number of trade-offs in terms of inference. The closer an experiment is to the field, the closer it is to the actual social-life phenomenon one seeks to study. This amounts to better *external validity*, a methodological challenge to be discussed in Chapter 8. At the same time, the aim to be closer to the field also implies a larger set of 'dimensions' at stake in the experiment, many of which cannot be controlled or randomised. This reduces the ability to provide 'pure' causal evidence, the laboratory becoming a natural place to test-bed causal measures before moving to situations closer to their naturally occurring environment.[11]

A third concern is related to the ability to actually randomise the treatment variable in order to produce the observations needed to estimate the causal effects. The discussion in Section 3.4.1 clearly shows that what one needs to achieve identification are two random samples of individuals: one in which all individuals receive the treatment, one in which nobody does. Two phenomena may typically challenge this requirement in field experiments. The first is *attrition*, i.e. when, for any reason, some subjects drop out of the experiment. If such drop out is purely random, then the consequences are only a matter of sample size as some observations will no longer be available. But a more serious concern arises if dropouts are specific to individuals who carry some specific heterogeneity related to the outcome variable. The dropout is then endogenous, and identification is challenged despite the random assignment. It will typically be the case if people who give up do so because they realise they cannot expect much improvement in their situation thanks to the programme. The second kind of phenomenon restricting the scope of random assignment in the field is *non-compliance*. This refers to the fact that subjects do not behave according to the rules or framework set out in the experiment. This may very well happen, as a field experiment takes place in reality, and the ability to 'constrain' people to follow rules is naturally limited. In terms of identification, it means that some people in the treatment group actually did not receive the treatment, or received a treatment that is different from the one studied. In both cases, the experiment does not deliver the observations on outcomes required to identify the effect of interest (see Deaton, 2010; Greenberg and Barnow, 2014, for more detailed discussions of such limitations to the identification achieved by field experiments).

A last concern is that treatments in the field may have spillover effects: because the change occurs in real life, changes in the situation of treated individuals may have an impact on untreated ones. Crépon et al. (2013), for instance, show that the enhanced probability to be employed of job seekers who benefit from a job placement programme actually comes at the expense of a lower probability of employment among similar job seekers who did not benefit from the programme. A simple comparison in outcomes between the two groups provides a highly biased measure of the treatment, as this effect magnifies the difference in outcomes between individuals from the two groups.

All the limitations listed above, and the many others discussed in the literature, can be addressed by either carefully designed experiments or additional experimental evidence

[11] Al-Ubaydli and List (2015) advocate an opposite perspective on this point, based on the ability of field experiments to bypass participant self-selection.

from the field. The main lesson from this comparison is actually not to name the one experimental environment that is best suited to serve research questions. The comparison rather shows the strengths and weaknesses of the two methods. To sum up: laboratory experiments provide strong control over the data-generating process, hence making the measure of causal effects easier and more convincing – how it is achieved is the purpose of Chapter 5. This generally comes at the price of a highly artificial decision environment, leading to strong doubts on their ability to describe what happens in real life. Field experiments, by contrast, provide evidence in their naturally occurring context by running experiments inside 'real life' itself. In this much more realistic context, however, many aspects of the environment are beyond the control of the experimenter, resulting in weaker inference properties.

Summary

Inference based on empirical observations involves two different issues: statistical inference, the relationships between sample observations and their population equivalent, and identification, the ability to measure a well-defined parameter based on available observations. Econometrics adjusts data analysis methods to the specific data and inference issues under study thanks to the elicitation of the data-generating process. Whether the assumed data-generating process in the econometric model matches the properties of the true data-generating process of the data is the key driving force to the ability to achieve proper inference. As regards identification issues, this consistency leads to identifying assumptions – properties of the data-generating process that condition the identification properties of different kinds of estimator. When the econometric model is linear, for instance, the identifying assumption is that the noise in the outcome variable is not correlated with the covariates.

The policy evaluation literature applies these ideas to the identification of a causal effect induced by a spontaneous change in the economy, called natural experiments. This literature provides a discussion of the true parameters of interest in such situations, and identifying assumptions that can be used to estimate them. The average treatment effect (ATE) is the expected causal effect of the change on a randomly drawn individual from the population. The average treatment on the treated (ATT) is the expected causal effect of the change on those individuals who actually face it. In estimating any of these parameters, the challenge is to find counterfactual observations – i.e. observations of what would have happened in situations that are not observed in the data.

Actual experiments differ from other empirical approaches in that they provide control over the data-generating process itself, by means of the design of the experiment and the choice of experimental treatments. Identifying assumptions in econometrics, and more precisely in the policy evaluation literature, thus provide a framework on how to design experiments to achieve proper identification and statistical inference. This leads to several strategies: randomisation, which allows us to break any possible correlation between what is unobserved (the noise) and the covariates of interest; measurement, which develops tools to include in the list of unobservable covariates that could be

confounding; and control, which holds constant any source of variation that is possibly confounding. This ability to choose the data-generating process is shared by all kinds of actual experiments along the continuum defined by the many dimensions in which experiments can be close or far from the field situation under study. The choice over the intensity of the 'fieldness' of an experiment implies a trade-off that balances (i) the improved identification offered by the strength of control over the environment in laboratory experiments and (ii) the generalisability of the results when they are observed in their naturally occurring environment.

4 The Need for Experimental Methods in Economic Science

As discussed in Chapter 3, experiments answer a specific need in empirical economics thanks to their ability to provide proper identification of the true parameters embedded in the data-generating process. The nature of such parameters, i.e. the reason why researchers are interested in their empirical identification, still remains an open question. This question is important for a critical assessment of the quality of identification: as shown in Chapter 3, empirical identification always hinges on (untestable) assumptions about the true data-generating process. The set of assumptions achieving identification, on which the interpretation of the results relies, heavily depends on the aim of the empirical study.

In order to answer such questions, this chapter first reviews the commonly recognised objectives of laboratory experiments – testing theory, searching for facts, or informing public policies. Each of the three involves a particular kind of dialogue between economic experiments on the one hand and economic theory and/or reality, on the other hand. To understand how laboratory experiments can serve these goals, it is necessary to clarify the relationships between experiments, theory and reality. These definitions will make clear that each aim comes with specific identification assumptions conditioning the usefulness of the estimated parameters. When testing theory, the first-order question is whether or not the observed behaviour is induced by the decision environment replicating the theoretical model – an issue defined as internal validity in Chapter 3. When searching for facts, by contrast, the main question is whether or not the same behaviour is to be observed outside the laboratory environment – a matter of external validity; while, finally, informing public policies requires both theoretical insights and well-documented facts, so that both internal and external validity are required. This discussion will thus provide the main background of the topic of the remaining parts: how the design of an experiment is adjusted to achieve its goal, and what experiments accordingly tell us about the specific research question stimulating their implementation.

4.1 What Laboratory Experiments Aim For

As all social sciences, economics aims to provide an understanding of social reality – how people behave and why? Can we collectively do better than what we currently do? And if yes, how? Unlike most social sciences, though, the main tool used in economic science to that end is to build formal theoretical models. Because it is both empirical and

highly controlled, laboratory experiments stand in the middle of this continuum between theory and reality. According to the seminal classification of Roth (1988), the three purposes that laboratory experiments are able to serve actually puts a bridge between theory and reality in both possible directions. First, experiments can be used for *testing theory*, i.e. assessing the empirical relevance of theoretical models. Second, they can be used to *search for facts*, in which case they use reality to inform theory – experiments then being used as 'exhibits' rather than 'tests' (Sugden, 2005). Last, building on both kinds of contribution, experiments can be used for *whispering in the ears of princes*, i.e. informing the decisions of policy-makers.

Testing theory

Theoretical models rely on behavioural assumptions to provide an understanding of the decisions of agents, and the resulting outcomes, induced by a given environment. They do so by restricting the economically relevant situation to a few key features. The Vickrey auction experiment reviewed in Chapter 2 is an example of such an exercise. Vickrey auction models reduce the auction environment to marginal values for the good, prices and monetary benefits. Based on the assumptions of utility maximisation and the axiomatic underlying the game-theoretic analysis of strategic interactions, it yields clear-cut predictions of both bidding behaviour and the properties of the resulting allocation.

As illustrated in Chapter 2, experiments exhibit two major advantages in that regard: the ability to both build an empirical situation that mimics the theoretical model, and measure or observe usually non-observable, or hardly measurable, dimensions (such as individual preferences towards the good, or individual prices posted). Experiments also get theory closer to reality by providing measures of individual preferences. As will be illustrated in later applications (see e.g. Sections 5.6, 6.6 and 7.4), such procedures rely on decision environments in which observed behaviour provides a direct measure of individual intrinsic attitudes. It not only allows us to assess whether the theoretical account of preferences actually makes empirical sense (e.g. to what extent is behaviour in risky environments actually described by the assumptions of expected utility), but also to assess whether predicted behaviour based on such preferences coincides with what theory predicts. In all these instances, experiments help in assessing the empirical relevance of theoretical results in terms of accuracy, precision and extent.

Searching for facts

On the other side of the continuum, there are many economic situations that are worth understanding, but which are too complicated and/or too specific to be covered by theory. Auctions again provide a useful illustration of such experiments: as the allocation mechanism, or the amount of information available to bidders, becomes more specific, auction models quickly become intractable. This does not mean, of course, that such specific auction markets are of no economic interest (even when there is no obvious reason why a market works as it does, the mere fact that it is used in practice is sometimes enough to make it worth investigating). In such cases, experiments can be used as a substitute for theoretical analysis. They are used to 'search for facts' in the sense

that they allow us to mimic well-defined situations and measure behaviour as well as the outcomes they generate. In the absence of prior expectations based on theory, such observations provide empirical knowledge about how the environment works, and what are its most sensitive features. To serve this purpose, such evidence must be robust and conclusive enough to actually serve as a stylised fact. In that regard, the replicability of experimental data, and the possibility to assess the robustness of the results through variations of the environment, are important advantages of experiments.

Whispering in the ears of princes

This third purpose of experiments amounts to improving the decision-making process by informing regulators or decision-makers (Roth's 'princes') of the likely outcomes of new or existing public policies (see, e.g. Holt et al., 2006; Normann and Ricciuti, 2009, for surveys). The general principle is to use experiments to test-bed decision environments such as market mechanisms, policy changes or new organisational structures. Observed outcomes in the experiment provide insights into the likely changes in behaviour and economic outcomes raised by innovative decision environments. As such, this aim builds on the ability of experiments to both test theory and search for facts – depending on whether theoretical insights are available on the policy-relevant question under investigation. The specific contribution of experiments to policy design comes from the ability to answer the specific needs of decision-makers. Because all the parameters of the decision environment can be freely set in the laboratory, an experiment makes it possible to fully replicate the specific features of a given policy. Illustration 4.1 describes an early example of such a contribution. This ability to fine-tune the experimental environment according to the requirement of the policy-relevant questions stands in sharp contrast with observational data. As compared to field experiments, laboratory experiments are cheap and easily implementable. An additional contribution of the use of experiments to policy design is their use in an instructive function. Laboratory experiments make economic reasoning more intuitive and appealing for non-academics. Even without producing any new knowledge, they can be used to make a convincing case of what the consequences will be of an intended change in the environment.

4.2 Experiments, Theory and Reality: How Experiments Achieve Their Goals

The above-mentioned aims of laboratory experiments (testing theory, searching for facts and whispering in the ears of princes) all refer to some form of interaction between experiments on the one hand, and on the other hand either theory (first aim) or reality (second aim) or both (third aim). As a result, assessing the ability of experiments to achieve either of these goals requires an in-depth understanding of how '*reality*' and '*theory*' are defined from the point of view of economics.[1] We will discuss each of these separate elements in turn, before moving to a definition of laboratory experiments. This will allow us to clarify the interactions of experiments with theory and reality and characterise the main challenges they raise.

[1] This section relies on the framework and discussion presented in Samuelson (2005).

Illustration 4.1
Whispering in the ears of antitrust authorities

The 1979 *Ethyl Corporation* v. *Federal Trade Commission* case is one of the earliest examples of the use of experiments for policy design purpose (both the case and the experimental evidence are reported in Grether and Plott, 1984). This case opposed the FTC to producers of gasoline complements that were in widespread use at the time. The main concern of the FTC was about the possible anti-competitive effect of the contract practices used on this market. Three kinds of provision attracted attention:

- advanced notice and price announcement – any increase in price was announced at least 30 days in advance by producers;
- most-favoured-nation – producers commit to offer all buyers any discount subsequently offered to another;
- delivered pricing – producers post a list of price for a given compound, regardless of the location of the producer.

The FTC feared that such practices on a market with only a few producers might be used by firms as a means to maintain prices above their competitive level. But such an opinion was not grounded on either a rigorous theoretical analysis or any empirical knowledge. The FTC thus asked Grether and Plott to experimentally investigate this question. Based on experimental treatments replicating each of the above practices as well as various combinations of them, the authors show that they actually push prices upwards – as compared to the same competitive market without such provisions. This evidence subsequently stimulated a more formal analysis of the anti-competitive effects of such practices, showing that the theoretical evidence matches with the empirical strategies observed in the experiment (Holt and Scheffman, 1987). For the sake of completeness, it is worth noting that this experimental evidence was eventually not used as a piece of evidence during the judicial proceedings. However, it played a crucial role to reinforce the FTC in its will to open a case.

These definitions merit a few notations. We will denote $\mathcal{X}$ the inputs and $\mathcal{Y}$ the outcomes or consequences of a given situation (or environment). Both are defined as sets according to their K-dimensional combination:

$$\mathcal{X}^K = \prod_{k=1}^{K} \mathcal{X}_i \text{ and } \mathcal{Y}^K = \prod_{k=1}^{K} \mathcal{Y}_i$$

The set $\mathcal{X}^K$ thus denotes the combination of all the K inputs considered, and similarly $\mathcal{Y}^K$ is the set of all the K dimensions that are consequences of the situation.

4.2.1 What Is Reality?

Let's start with the most controversial definition: what is the real world? From the point of view of a scientist, who tries to understand why what happens happens, the answer is simpler than one would expect. In short, reality is no more than a set of causal relationships. This means that reality can be reduced to three elements: the causes $\mathcal{X}$, the consequences $\mathcal{Y}$ and a 'function', that transforms causes into

consequences. What makes reality a complex object, even in this overly simplified representation, is that inputs and outputs are infinity-dimensional. Therefore, reality is the function:[2]

$$F: \mathcal{X}^{\infty} \rightarrow \mathcal{Y}^{\infty}$$

In words, the function F transforms inputs into outputs. Each particular combination of the content of $\mathcal{X}^{\infty}$, that defines a situation of interest, causes a particular output from the set of $\mathcal{Y}^{\infty}$. Social science can to a large extent be reduced to an attempt to understand this function F: how causes and consequences are related together in real life. As straightforward as it may seem, this definition makes an important methodological point: it is impossible to understand reality as it is. The input and output dimensions are infinite, because reality features an endless set of potentially relevant properties and characteristics. As an example, the relationship between employers and their employees involves paying a wage and exerting an effort, but also a relationship between two persons, the consequences for their family life, relations with co-workers, etc. As a result, neither theoretical nor experimental research could ever describe all aspects of reality. In a sense, the complexity of the real world is the very reason why science is needed: if the world were easy to understand as it is, no one would need the help of a science to elicit what happens and why it happened.

4.2.2 What Is a Model?

Because reality is such a complex object, science in general and economics in particular proceed by breaking down the real world into only a handful of its components. Therefore a model is, and must be, a simplified account of a real situation. As such, a built-in property of a model is to be false, i.e. to be unable to fully account for all subtle drivers and consequences of a situation. This determines the main ingredient of a model.

Like the function F that maps causes to consequences in the real world, a model seeks to capture a particular causal relationship between a number of inputs $\mathcal{X}$ (causes) and outputs $\mathcal{Y}$ (consequences). Reducing the real-world situation of interest, so as to make it intelligible, also amounts to restricting the dimensions of the sets. Rather than all causes and all possible consequences pertaining to the situation, a model will only focus on a few causes and a few of their consequences. This choice is part of the model: focusing, for instance, on performance and compensation schemes, as economics does, is obviously a restrictive view of work situations. But as stressed above, the challenge is not to understand reality as it is, but rather to enlighten some aspects of it and reduce the main mechanisms to a few channels – so that it becomes possible to think about the situation from inside the model. Illustration 4.2, for instance, shows how the behaviour of complex human organisation like firms, acting in many different ways on markets, are

[2] The aim of this discussion is definitely not to provide a full account of the epistemology of (social) sciences. We ignore, in particular, the important and interesting question whether one can actually separate reality and the tools and prism used to analyse it.

Illustration 4.2
Models as a reduction of reality: firms' behaviour in collusion theory

Collusion theory in industrial organisation differs from standard competition in that firms' decisions are assumed to be strategic – typically because only a few of them compete on the same market. We denote N the number of firms. The market is reduced to only a few dimensions, essentially prices and quantities, and the firms' behaviour is described by a profit function. The main focus here is to understand the difference in behaviour induced by the horizon of the decision-making. When competing in a one-shot set-up, each firm has an incentive to beat its competitor's decisions – resulting in the competitive, zero-profit, equilibrium if firms compete in price and the Cournot–Nash equilibrium if they compete on quantities. Let y be the quantities produced (and sold) and q the market price, $\pi(q, y)$ denotes the profit firms receive at the one-shot equilibrium, and $\pi_{\mathrm{M}}(q, y)$ the one-shot monopoly profit. The important distinguishing feature of collusion is that firms are playing an indefinitely repeated game. Denoting δ the firms' exponential discount factor, collusion yields the following profit for each firm:

$$\sum_{t=0}^{\infty} \frac{\pi_{\mathrm{M}}(q, y)}{N} \delta^t = \frac{1}{1-\delta} \frac{\pi_{\mathrm{M}}(q, y)}{N}$$

Firms make this share of the monopoly profits as long as all firms collude and cooperate. As soon as one firm deviates from collusion, and cheats, all the other firms can use a (so-called) trigger punishment strategy, i.e. revert to the Cournot–Nash equilibrium from then on. The profit of a deviating firm is the sum of the one-shot monopoly profit $\pi_{\mathrm{M}}(q, y)$ the firm gets when it deviates, and of the discounted sum of the future Nash–Cournot profit $\pi(q, y)$:

$$\pi_{\mathrm{M}}(q, y) + \frac{\delta}{1-\delta} \frac{\pi(q, y)}{N}$$

Therefore, collusion is individually rational and can be sustained at equilibrium if no firm has an incentive to deviate:

$$\frac{1}{1-\delta} \frac{\pi_{\mathrm{M}}(q, y)}{N} \geq \pi_{\mathrm{M}}(q, y) + \frac{\delta}{1-\delta} \frac{\pi(q, y)}{N}$$

In the case of a Bertrand game in which firms compete in prices, the individual one-shot profit $\pi(q, y)$ is equal to 0. Therefore the collusion is sustainable if:

$$\delta \geq 1 - \frac{1}{N}$$

In the context of this model, collusion is a stable equilibrium (hence arising on actual markets) when the threat of punishment (through trigger strategies in this case) is large and credible enough. This is more likely to happen when the number of firms N is low and when δ is high, i.e. when firms only slightly discount the future. For future reference, it is worth noting that the model would write exactly the same if δ stood for the probability that the market survives for one more period at the end of each market period.

reduced to a few of its components in the analysis of collusive behaviour in industrial organisation.

This first step in theoretical analysis leads us to identify a few influential and/or interesting dimensions of the situation. On top of this finite number of inputs, denoted $n_{\mathcal{X}}$,

and a finite number $n_{\mathcal{Y}}$ of outputs, a model is again a causal relationship between the two:[3]

$$f : \mathcal{X}^{n_{\mathcal{X}}} \rightarrow \mathcal{Y}^{n_{\mathcal{Y}}}$$

The function f maps the set of inputs to the respective outcomes. Its aim is to be a 'sub-function' of F, hence to reflect part of the causal mechanisms at stake. The challenge in developing a model and assessing its quality is to solve the trade-off raised by the need to simplify reality. In the choice of $n_{\mathcal{X}}$ and $n_{\mathcal{Y}}$, the model must get rid of all unnecessary details, which would otherwise cloud the focus on the relationship between the causes and consequences of interest. But the theory cannot be too simplistic either, as it would then lose its explicative power, and its ability to be generalised. To sum up, without giving any definitive answer, a model must be as abstract as possible, and as detailed as necessary, in order to provide an accurate and closely focused account of reality.

4.2.3 What Is an Experiment?

An experiment is a controlled situation in which many features of the environment are implemented by design, so as to observe the resulting individual decisions and interactions. The aim of such observation is to infer the causal relationships between the environment and individual(s) behaviour. This can be easily embodied in the current notations: an experiment is nothing but a choice of a set of n inputs, defining the environment, associated with m measures of their consequences with the goal of inferring their causal relationship, F.

But, at the same time, an experiment deals with the decisions of human beings in a particular real-world environment – the laboratory. Thus, the measured outcomes in the laboratory are not only the results of the chosen controls, but also the consequence of an endless range of influences: from personal characteristics to anything the person may have experienced prior to coming to the laboratory, or any specificities of what happens inside and outside the laboratory during the experiment. An experiment is thus an almost-real situation, with many dimensions beyond the n inputs that are controlled and chosen actually influencing what happens. The situation built in the laboratory is thus defined by $x^{\infty} \in \mathcal{X}^{\infty}$, such that $x^{\infty} = x^n \cup x^{\infty - n}$, i.e. an element $\mathcal{X}^{\infty}$ that matches x^n (only) on the first n dimensions but leaves the others uncontrolled. Typical examples of such uncontrolled inputs are the level of noise inside and outside the laboratory, the mood of participants when arriving at the experiment, etc. These are all factors that possibly influence the behavioural response to the experimental situation, without belonging to the components that were chosen when designing it. In this setting, the outcome from an experiment is a set of measures m such that:

$$F^m(x^{\infty})$$

This outcome must be thought of as a subset of the infinite-dimensional function F describing all the consequences of an experiment. Again, while individual decisions and payoffs will typically belong to the set of m measures that will be observed, there

[3] In order to better fit models that allow for heterogeneity, the definition can be easily generalised, as in Samuelson (2005), to a distribution over consequences rather than a deterministic relationship.

will always remain many consequences of the inputs of the situation that will not be observed: how the mood of participants changed during the experiment, what is the induced change in the room temperature, etc.

Two important lessons can be drawn from this definition. First, a critical aspect of the design of an experiment lies in the choice of the finite number of inputs, n, and the finite number of measures, m, on which the design is based – above and beyond the actual implementation of the input controls and the accurate observation of the outcomes of interest. The choice depends entirely on the relationship and phenomenon that the experiment seeks to measure – i.e. the research question. The case study presented in Section 4.3 illustrates this idea.

Second, this definition makes it clear that experiments share common features with both theoretical models and real economic life, as defined above. Like a model, an experiment focuses on a subset of relevant dimensions in order to study the phenomenon of interest. The aim is not to provide observations on all consequences induced by all relevant factors at stake in the empirical situation of reference, but rather to pin down a few driving forces and a few consequences, on which empirical evidence is needed. But at the same time, like in reality, what happens in the laboratory is made of an infinity of causes and consequences which no one is able to control or observe. Even with cleverly designed experiments, there will always be an infinity of inputs that will remain uncontrolled.

Because experiments have many of the same features as both theoretical models and reality, they can be seen as a way to establish a bilateral link between theory and the social reality this theory aims to understand (Croson and Gächter, 2010). This is the main reason why experiments are a particularly effective way of enriching and deepening, but also complementing or challenging, our understanding of the real world through economic theory and economic analysis. This is, in particular, the reason why they are well suited to achieve either of the goals described earlier. How they achieve it is derived from their attractive empirical properties in terms of identification, as described in length in Chapter 3. On this basis, experiments allow us to observe the consequences of the inputs involved – i.e. $F^m(x^\infty)$ – or the variation induced by different sets of inputs – through the comparison of $F^m(x^\infty)$ and $F^m(x'^\infty)$, i.e. the differences in outcomes under two sets of inputs x and x', which define treatment effects. Depending on the aim of the experiment, the inputs and measures will be defined according to either an underlying theory to be tested or a real-world phenomenon one seeks to understand better. The condition under which the outcome will be informative, however, strongly depends on the specific goal of the experiment. The remainder of the chapter considers each of them in turn.

4.3 *Case Study*: Deepening Understanding through Additional Controls and Measures: The Dictator Game

This section aims to illustrate how the choice of inputs and measures is part of the design of an experiment, and how it allows us to incrementally enrich the conclusions drawn

from experimental observations. To that end, we focus on one of the most replicated game in the experimental economics literature: the dictator game (DG).

In this simple game, first introduced by Forsythe et al. (1994),[4] two players interact in a strictly anonymous setting. Both players do not know who their partner is, nor do they find out at any point. Within a pair of two subjects, one person is assigned to be the 'dictator', and the other one the 'receiver'. The dictator receives an endowment, e_i, and makes an offer $e_i \geq \ell_i \geq 0$ to the receiver. As a result, the dictator gets $e_i - \ell_i$ and the receiver gets the offer ℓ_i. There is no constraint on the split decided by the dictator. The open question in this environment is: how much money is the dictator willing to give to the receiver? Economic theory is of little help to answer the question – a point the experiment has been designed to make, as a matter of fact. Since there is neither any monetary benefit, nor any constraint on the amount given to the receiver, the dictator should 'theoretically' keep as much money as possible for him-/herself. This is obviously not what happens when people are asked to make such decisions in an experiment implementing the game.

Figure 4.1 provides an overview of the patterns of behaviour generally observed in this kind of experiment, based on a meta-analysis including more than 300 published studies all replicating the dictator game (Engel, 2011). While it comes as no surprise that some people do donate something to their receiver, the extent to which they do so is puzzling. On average, the dictators give 28.35% to a complete stranger whom they will never meet. It is not only the magnitude of the donation that sparked a great deal of interest in the literature, but also the strong heterogeneity in donation behaviour. One-third of people behave in accordance with a purely self-interested model of decision-making, one-sixth offer an equal split and one out of 20 even offers everything. These patterns are generally seen as robust stylised facts, thanks to the many replications of the same game with slight variations in the design (in different countries, with varying amounts of money to be split, etc.). In order to better understand the reason why such behaviour occurs, what are the motivations behind it and what fosters or undermines it, additional inputs and outputs have been considered in the literature.

4.3.1 Additional Inputs 1 – Social Distance

A first hypothesis is that donation behaviour is related to social relationships. In order to investigate this point, Hoffman et al. (1996) conducted six different versions of the dictator game, varying the dictator's degree of distance, isolation and anonymity – while maintaining the same decision-making structure.

The first treatment (denoted FHSS-R) is an exact replication of the Forsythe et al. (1994) experiment, the amount of the dictator's initial endowment being $e_i = \$10$. FHSS-V is a replication of the same game but uses neutral wording for what the game is about. The researchers did not use the 'sharing' language of the first version and instead asked the subject to 'divide' the preliminary allocation. Two further treatments strengthen the social distance involved in the dictator's decision-making by

[4] The Forsythe et al. (1994) game is itself a simplified version of a binary decision game introduced by Kahneman et al. (1986).

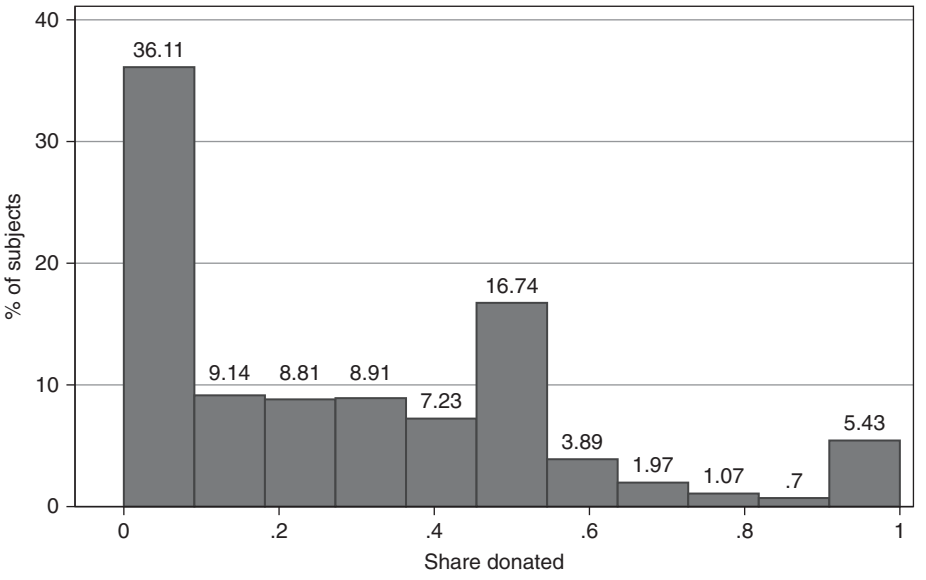


Figure 4.1 Meta-analysis results: the dictator game

Note. Empirical distribution of the population of dictators observed in more than 300 published studies, according to the share of the initial endowment given by the dictator to the receiver.
Source: Engel (2011, p. 589, Figure 2).

implementing a double-blind donation. The procedures are drastically altered to achieve such isolation. First, the experiment is run by a subject, randomly selected to act as a monitor and paid to do so. But the most important changes are related to the elicitation of the donation decision. In an experiment involving 14 subjects playing as dictators, each subject draws an envelope from an urn. It is common knowledge that 12 of the envelopes in this urn contain 10 one dollar bills and 10 blank slips of paper, while the remaining two are filled with 20 blank slips of paper. Dictators then go to a separate room, keep as many bills as they wish and replace the bills with blank slips of paper. They seal the envelope, bring it back in a box, and leave the room. Once all dictators have made their decisions, the 14 subjects playing as receivers come one by one to draw an envelope from the box, show the content to the monitor and leave the room with it.

Thanks to the two blank envelopes, the procedures in this DB1 treatment imply that neither the experimenter nor the receiver is able to identify the individual decision of the dictators. Since it is common knowledge, it is expected to increase the social isolation of the dictators' donation decision. A second variation, DB2, weakens the anonymity achieved in this treatment by removing the paid monitor and the two blank envelopes – it is still the case that individual decisions are unobserved by both the receiver and the experimenter, but the distribution of decisions is now perfectly observed. Two additional treatments are meant to disentangle the components of social isolation induced by these treatments, by implementing a single-blind donation decision. In SB1, the experimenter opens up the envelope when the dictator brings it back, hence breaking

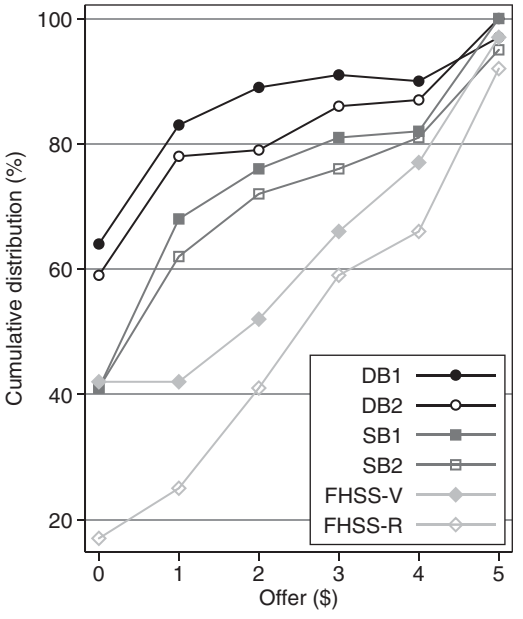


Figure 4.2 The effect of social distance on dictators' decisions

Note. Cumulative distribution of the amount donated by dictators in each treatment.
Source: Hoffman et al. (1996, p. 654, Figure 1).

the anonymity with the experimenter. Last, in SB2, the envelope into which the dictator puts the amount is replaced by a transaction form, which the dictator has to fill out and bring to the experimenter in an envelope so as to be paid the corresponding amount of money. The experimenter opens the envelope and notes the subject's name and decision and gives the dictator the sum $e_i - \ell_i$ they decided to keep. This procedure decreases the relative social distance, as the subject enters into a real transaction with the experimenter.

Figure 4.2 shows the cumulative distribution of donated amounts observed in the experiment (in this experiment, donations are never higher than five dollars, the equal split): the steeper the line, the less dictators keep for themselves. As compared to the baseline, the change in wording implemented in FHSS-V slightly decreases donations. But the most impressive change in behaviour comes with increased social isolation: there is a perfect first-order stochastic dominance relation in the distributions of dictators' behaviour according to the blindness of the donation decisions. In DB1, where the level of anonymity is the highest by design, only 40% of the dictators leave positive amounts to the receivers.

This experiment thus provides strong evidence that donation behaviour in dictator games is sensitive to social image. This is done by adding to the list an input, social distance, that wasn't part of the design of the baseline game. It is worth noting that such contrasting evidence between the two series of experiments does not dismiss any of the two results: surprisingly high donations in the baseline are an empirical fact, just as is the reduced donation when social distance is high. The actual, and more interesting,

question is rather which one of the two kinds of behaviour is more relevant, and what each one of them tells us about the underlying reason behind the other.[5]

4.3.2 Additional Inputs 2: Earned money

Cherry et al. (2002) focus on another dimension of the original game: the property rights over the amount of money to be split by the dictator. In the baseline experiment, the endowment is 'windfall' as it comes to the participants from out of nowhere. Cherry et al. (2002) analyse the effect of changing windfall money into earned money.

This is achieved through a preliminary task, in which subjects earn their endowment according to a performance-based payoff. The issue with such an experimental treatment is that the size of the amount of money to be split can change donation behaviour by itself, beyond any effect of the property rights over this amount. To address this issue, the earned-money treatment relies on a tournament based on the performance at the preliminary task: those subjects whose performance is above the median are given \$40, while other subjects are given \$10. This treatment thus elicits observation on donation behaviour with earned money over two possible values of the endowment: high (EH treatment), or low (EL treatment). In order to measure the effect of windfall money on donation behaviour, two versions of the baseline treatment with windfall money are considered: BH (the windfall endowment is set equal to \$40) and BL (the windfall endowment is set equal to \$10). A third treatment couples the earned-money preliminary task and the double-blind treatment introduced by Hoffman et al. (1996), generating again observations of donations under two possible values of the endowment: DBH and DBL.

Observed donation behaviours in each treatment are organised in Figure 4.3 according to the amount of the endowment. In the baseline, cumulative distributions of donation behaviours again display strong heterogeneity, with 20% of the sample giving nothing to the receiver and the remaining offering up to half the endowment. Earned money drastically affects donation behaviour in both dimensions. The number of subjects who decide to give nothing in this case is three times higher, and positive amounts not only appear more often but also are of lower magnitude. When coupled with double blind procedures, earned money almost perfectly 'hardnoses' the dictator – only a few subjects still send small, but positive, amounts to receivers. These results unambiguously confirm that the legitimacy of the endowment plays a crucial role in shaping other-regarding behaviour. Again, the aim here is not to disqualify any of the two kinds of observation, but rather to characterise the dimensions of the environment that most crucially determine behaviour. Thanks to additional inputs in the experimental design, this experiment underlines the influence of the nature of the endowment.

[5] In the words of Hoffman et al. (1996, p. 654), 'this experimental exercise is fundamental to understanding the received evidence for other-regarding behaviour that is frequently manifest in bargaining game experiments, but in which strategic reciprocity and utilitarian elements are confounded in interpreting observed outcomes'.

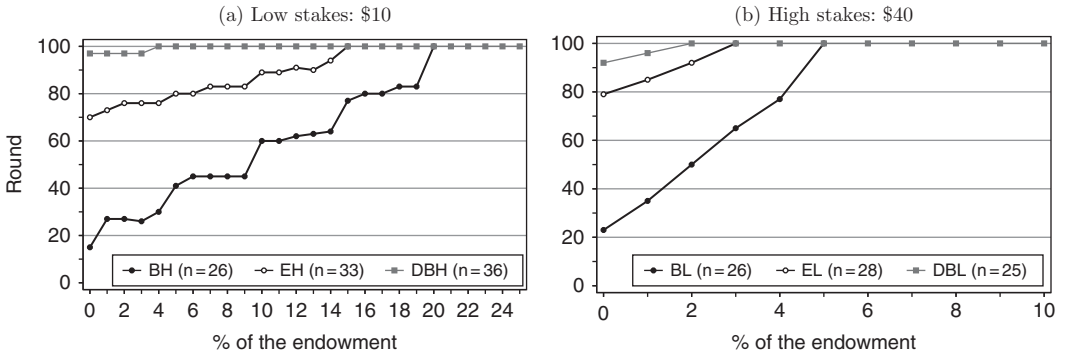


Figure 4.3 Offers in the dictator game with earned money

Note. Cumulative distributions of donation behaviours out of $10 in (a) or $40 in (b), endowment in the baseline (BH, BL), earned-money (EH, EL) and double-blind (DBH, DBL) treatments.
Source: Cherry et al. (2002, p. 1220, Figures 1 and 2).

4.3.3 Additional Inputs 3: Property rights on Player Positions

Beyond the nature of the endowment, another source of concern in the game is the strong asymmetry of positions between the dictator and the receiver: one subject is given full power over the outcome, while another is given no choice but to passively experience the decisions made by another person. Hoffman et al. (1994) study this dimension of the decision-making problem by introducing an additional treatment with earned entitlement to behave as a dictator. In a preliminary stage, all subjects are asked to answer a quiz on current events. After the quiz, subjects are ranked and then split into two groups: the top half of the subjects become dictators and the bottom half become receivers. Lastly, a matching procedure pairs the top-ranked dictator with the best receiver, the second-best dictator with the second-best receiver, and so forth.

Figure 4.4. shows the results of the experiment, along with those from the baseline treatment corresponding to the double-blind dictator game with windfall money as described in Hoffman et al. (1996; see Section 4.3.1 above). It clearly appears that the amounts shared are considerably lower when there is an earning stage which precedes this decision – one out of 10 subjects offers an equal split in the baseline treatment, while only 5% of subjects offer $4 out of $10 in the earned entitlement treatment. Property rights over the position in the game, which are ruled out in standard implementations of the dictator games by randomly allocating subjects to roles, thus foster selfishness.

4.3.4 Additional Measures: Response Times

Incrementally widening the set of inputs that are actually controlled in the experiment allows us to build more and more precise answers to research questions, by testing alternative hypotheses about the driving forces of behaviour. It is somehow too often neglected, but the exact same logic also applies to experimental outcomes. Among the many things that arise in the laboratory, and might be affected by the inputs involved, only a small subset is actually recorded so as to be part of the measures available. The

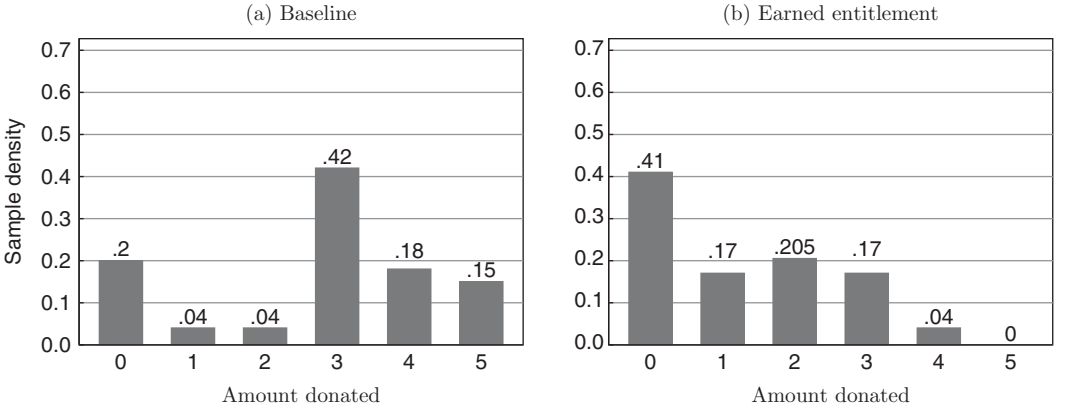


Figure 4.4 Donations from dictators who earned their position

Note. Empirical distribution of the amount donated by the dictator (in abscissa, out of an initial endowment of $10) in the control and earned-entitlement conditions.
Source: Hoffman et al. (1994, p. 365, Figure 4).

empirical answers generated by experiments can be very different depending on the pre-defined choice of measures. A typical example is the use of physiological measurement devices, such as skin conductance or eye tracking (see e.g. Sanfey, 2007, for a survey of applications to the dictator game), or the use of neurological measures (such as an fMRI), to investigate the brain activity while decisions occur. They all provide additional measures of the outcome from the experiment, that would always exist, but would have remained ignored, absent their implementation. Another example is response times.

The decision-making literature tends to relate different motives behind behaviour to the time spent on decision-making – as discussed in Focus 4.1. This can be easily investigated in the laboratory context, by keeping track of time elapsed from the beginning to the end of the decision task. Because it is simple and non-strategic, the dictator game is one of the first to which such measures have been applied. The results remain rather mixed. For example, the preliminary results obtained by Rubinstein (2006b, not reported in the 2007 published version of the paper) suggest that egotistic decisions are taken more slowly. On the other hand, the results obtained by Piovesan and Wengstrom (2009), and shown in Figure 4.5

Focus 4.1

On the use of response times to interpret observed behaviour in experiments

There is a growing interest in the correlation between economic behaviour and response times in both economics (e.g. Rand et al., 2012; Schotter and Trevino, 2012) and experimental psychology (Shalvi et al., 2012; Gino and Mogilner, 2014, among others). One of the psychological foundations of this focus on response times is based on the System 1/System 2 hypothesis raised by Kahneman (2003) – also known as the 'thinking-fast-and-slow' hypothesis (Kahneman, 2011). The model highlights a dichotomy between two modes of thought, System 1 being instinctive and emotional and System 2 more deliberative and logical. The main behavioural insight is that 'choices made instinctively, that is, on the basis of an emotional response, require less response time than choices that require the use of cognitive reasoning' (cited in Rubinstein, 2007, p. 1243). This pattern produces a strong correlation between mistakes and short decision times in decision problems where the solution is unambiguous. This leads Rubinstein (2007) to classify behaviour based on response times according to three types of action:

1. *cognitive*, for actions which involve a reasoning process;
2. *instinctive*, for actions which involves instinct;
3. *reasonless*, for actions which are likely to be the outcome of a random process with little or no reasoning about the decision problem.

Rubinstein (2013) collected the data on response times in a large-scale set of didactic online experiments. The evidence shows that the relationship between response time and error rates varies across tasks. For tasks where subjects can use a simple heuristic to avoid errors, response times are negatively correlated with error. This happens, for example, when participants have to choose repeatedly among alternatives. On the other hand, for tasks where the answers require a cognitive effort (coding letters, avoiding first-order stochastic dominance, assessing likelihood), response times are positive when correlated with errors. In the former case, simple heuristics benefit consistency. In the latter case, simple heuristics lead to mistakes. Lastly, and perhaps more importantly, the usual deviations from the expected utility observed in decision experiments (the Allais paradox, the three-colour Ellsberg paradox, Kahneman and Tversky's Asian disease problem) are not correlated with a given pattern of response times. The certainty effect, ambiguity aversion or framing effects are all compatible with fast and slow observed response times. For instance, Evans et al. (2015) suggest that response times might be related to decision conflicts rather than to dual thinking. Using a meta-analysis of existing experimental evidence, Rand (2016) confirms that favouring deliberation over intuition tends to push behaviour towards less pure cooperation, but leaves strategic cooperation unchanged.

when they enter the donation part of a binary dictator game – dictators are restricted to choosing between two possible allocations. During this task, subjects hear the sound of letters in headphones, and must press the corresponding letter on their keyboard but only for some of these letters – so that they need to stay focused and cannot just routinely press the key for all letters they hear. This treatment causes ego-depletion and lowers the ability to use deliberative decision as compared to the baseline, with no preliminary task. This cognitive load is applied to several conditions, in which the level

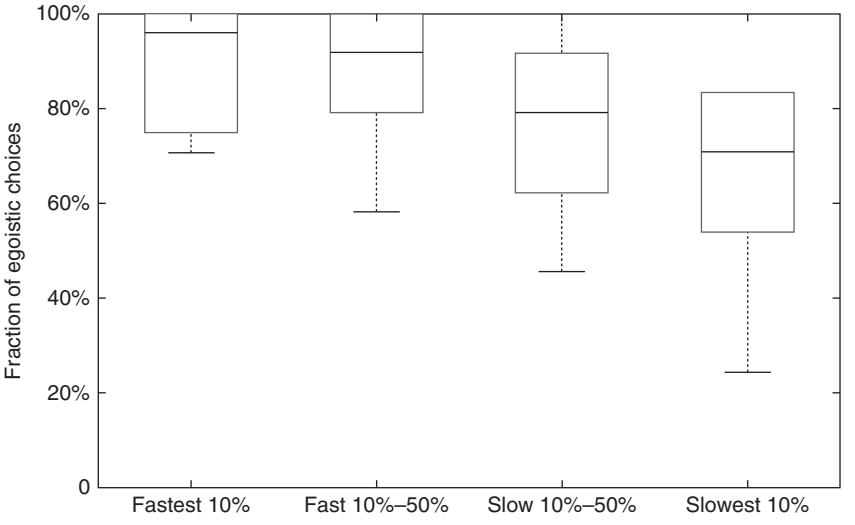


Figure 4.5 Generous decisions by dictators are taken slowly

Note. The figure reports the share of subjects who decide to behave selfishly in the dictator game in each subgroup defined by their quantile of response time (in abscissa). For each group, the straight line shows the median behaviour, the boxes cover the quartiles of the egoistic decisions distribution, and the vertical lines show the adjacent values.
Source: (Piovesan and Wengstrom, 2009, p. 195, Figure 2)

of inequality implemented by the two donation choices available to dictators is varied. The results show that lower deliberation leads to higher amounts of donations, and that deliberation leads to much more adjustment in the level of donation to the inequality of the allocation.

4.4 How Experiments Interact with Theory: Testing Models

Experiments share common features with both theory and reality, making it natural to use them to test-bed the empirical performance of theoretical models. The first challenge is to define an empirical environment that replicates the model assumptions. The interpretation of the results is heavily conditioned by the quality of the inference performed based on experimental data, i.e. the internal validity of the experiment.

4.4.1 Testing Theory

Theoretical models are most often way too general to be directly implemented: the actual decision-making environment must be simple enough to be described to people with no background in economics. The fair-wage-effort model described in Illustration 4.3 provides an example: the aim of the model is to underline the driving forces of a causal mechanism. For this purpose, the more general assumptions are, the more convincing and relevant is the model. But at the same time, such generality makes a theory compatible with many different specifications and concrete situations. The reason why

Illustration 4.3
Reciprocity at work: the fair-wage-effort hypothesis

The incentive effect of compensation schemes, illustrated all through Chapter 3, only goes through the change in consumption it offers to employees. Akerlof (1982) offers a 'gift exchange' model, the aim of which is to highlight the effect of norms and social relationships at work. The assumptions of the model are fairly simple. The worker's utility depends, as usual, negatively on the effort exerted on the job, ℓ_i, and positively on the wage w. But it also depends on a norm, ℓ_{norm}, which stands for the level of effort the worker sees as 'normal' given the work condition, and in particular the wage offered. The workers' utility function thus writes $u(\ell_{norm}, \ell_i, w)$. As is standard, a worker chooses the level of effort that maximises utility. The worker's choice of the utility-maximising level of effort is constrained by the minimum level of effort ℓ_{min} the firm requires – and implements through e.g. a control policy. The output of the firm depends on the work effort of all of its workers. The firm chooses the work rule ℓ_{min}, the wage function $w(\ell)$ and the number of workers it wishes to hire in order to maximise profit. The firm's behaviour is constrained by the worker participation constraint and takes the norm ℓ_{norm} as given. Akerlof (1982) considers several possible versions of labour market models based on this set of assumptions. They all illustrate the same main result: gift exchange at work as an incentive mechanism arising in equilibrium. One can assume, for instance, that all workers are homogeneous and exert an effort equal to the norm; and that this effort norm is a function of the firm wage relative to a reference wage w_0: $\ell_{norm} = -a + b(w/w_0)^{\gamma}$, with $\gamma < 1, a > 0, b > 0$. Akerlof (1982) defines the reference wage, w_0, as the geometric mean of the outside wage and the unemployment benefits. A direct consequence of this set of assumptions is that firms that pay a wage above the reference wage move the effort norm up and obtain extra effort. Firms thus optimally offer wages that are higher than w_0 – and fixed wages become incentive-compatible by inducing higher effort even if they do not directly link consumption and effort at work. This fair-wage mechanism is based on a 'gift-exchange' principle – a higher wage is considered a 'gift' by the employee, which reciprocates with higher levels of effort, to the benefit of the employer. Contrary to alternative models of efficiency wages (e.g. Shapiro and Stiglitz, 1984 in particular), the level of wage itself induces higher effort: not only will any wage cut reduce the effort of all the workers, but any positive wage shift will increase the effort of all the workers.

experiments nonetheless are an appropriate tool for testing theories is because general theories must also apply to simple cases they embed. In the words of Plott (1991, p. 902): 'General theories must apply to simple special cases. The laboratory technology can be used to create simple (but real) economies. These economies can then be used to test and evaluate the predictive capability of the general theories when they are applied to special cases'.[6] Experiments offer a simple way to create such simple situations. By doing so, they provide an empirical counterpart to theoretical models: they allow us to compare the empirical distribution of the behaviour obtained with the distribution predicted by theory; or to contrast the variation in empirical outcomes induced by an experimental treatment to the comparative statics generated by the model.

The gift exchange game described in Illustration 4.4 offers an example of such a process. It aims to provide an experimental implementation of the Akerlof fair-wage

[6] Also see Plott (1982; 1989) for earlier insightful discussions of the use of experiments to test theories.

Illustration 4.4
Experimental evidence on the fair-wage-effort hypothesis

Fehr et al. (1993) consider a very simple case of the Akerlof (1982) model in order to test its main lesson: that gift-exchange motives induce positive wage–performance relationships, so that high wages arise in equilibrium. Subjects are split into two groups: employees and employers, who play together a two-stage game. The first stage is a contracting game played in real time. Employers make increasing wage offers and employees either accept or reject the offers. Once an offer has been accepted, the employee is matched to the employer so that only unmatched subjects remain on the market. In the second stage, workers privately choose an effort level and receive the wage agreed upon in the first stage, whatever their effort decision. This wage is a cost for the employer, and a benefit for the employee. The payoffs are moreover designed in such a way that higher effort reduces the workers' payoffs and increases the employer's earnings. More precisely, the effort cost function $c(\ell_i)$ is a convex function defined from $\ell_i \in [0.1; 1]$ to $c(\ell_i) \in [0, 18]$. The payoff of worker i choosing effort ℓ_i and receiving wage w_j from employer j is $u_i = w_j - c(\ell_i) - a$, where a (set to 26) is a fixed cost of accepting a wage contract. On the firm side, the payoff of employer j is set to $\pi_j = (q - w_j)\ell_i$, where q is the unitary return to work (set to 126), and the wage cost is made proportional to effort of worker i so as to avoid losses (which reinforces the conflict of interest between employers and employees). If all subjects are pure payoff maximisers, the incentive structure of the experiment gives rise to clear-cut predictions: there is no incentive for workers to choose any level of effort higher than the minimum level, set equal to 0.1. This is further reinforced by the structure of the markets. In all experimental sessions, there is by design an excess supply of labour: the market always gathers more workers than employers. The ratio of excess labour supply is 9 workers for 6 employers in 3 out of the 4 sessions, and set to 8/5 in the fourth session. This, by itself, should push wages down by giving the whole bargaining power to employers. The first stage formally mimics a one-sided oral auction, which is known to converge theoretically and empirically to the competitive market price. Since the opportunity cost of accepting an offer is equal to 26, the competitive wage in the experiment is equal to 30 (since wages are chosen by step of five by design). Anticipating that no payoff-maximising worker would choose a level of effort higher than the minimum, this is the maximum level of wage employers should offer. The main results observed in the experiment are presented in the table below (from Fehr et al., 1993, p. 446, Table 2).

Wage	Average observed effort level	Median observed effort level
30–44	0.17	0.10
45–59	0.18	0.20
60–74	0.34	0.40
75–89	0.45	0.40
90–110	0.52	0.50

The results confirm the two main insights from the Akerlof (1982) model. First, the behavioural assumption that workers positively react to high wages through increased effort is confirmed – by looking at both the average and the median effort. Second, firms actually expect such an effect: despite the strong market forces pushing wages downwards, high levels of wages (up to almost four times the competitive level) are actually observed.

model. The experiment is designed in a such a way that the environment complies with the model's main assumptions – a work relationship with conflicting interests of the employer and the employee, competitive pressure on the choice of wage, non-contractible effort level, non-credible promises to exert high effort in exchange for a high level of wage due to the sequentiality of decisions. This environment is built in a way that is intuitive and simple enough to be easy to understand, and credible for subjects participating in the experiment. Observed behaviour can then be used to assess the validity of the model's main mechanisms and its predictions.

A testing process of this type is often seen as a simple accept-or-reject decision rule: theory is either confirmed, or challenged, depending on whether observed behaviour complies with the model predictions. As will be described in detail in Chapter 9, this question in itself is far from trivial. In a nutshell, this comes from three main reasons. First, no theory aims to perfectly predict behaviour, even in the simple world that directly stems from its assumptions. Deciding what is acceptable noise, and what observation definitely contradicts theoretical expectations, is often a difficult task. Second, the same kinds of theory often receive very different empirical assessments. As a simple example, double-auction theory (as studied by Smith, 1962, and described in Section 1.2.1) is generally seen as 'working': the induced behaviour of subjects in a double-auction, experiment in which both the buyers and sellers submit their bids, yields the same results as expected by economic theory. In contrast, the Nash equilibrium prediction in the prisoners' dilemma only holds to some extent in experiments (see Section 1.3.1). In both cases, the Nash equilibrium combining rationality assumptions and pure self-regarding preferences is the theoretical tool used to generate models' predictions. Testing models in such contexts often helps better characterise the situations to which theory applies and those in which it fails, rather than simply confirming or disparaging it.

But, perhaps more importantly, testing theory also means more than just seeking to accept or reject decisions. First, experiments can be used to disentangle competing models. Such an exercise is often hard to implement based on observational data, as competing models might rely on subtle assumption differences about the environment or economic agent's preferences. Experiments, by contrast, can be designed as simple cases in which observed outcomes can be contrasted with testable restrictions from each of the competing models. Second, an important role of testing theory is also to assess the empirical content and extent of theoretical assumptions (what Schram, 2005, labels 'stress tests of the theory'). Unrealistic assumptions abound in theoretical approaches. They are part of the process of simplifying and reducing real-world situations to get rid (deliberately) of part of this reality. To give just one example, atomicity on markets (leading to the important consequence that economic agents behave as price takers) is made of meaningless assumptions from an empirical point of view. The important and interesting question, however, is not that much whether each and every one of these assumptions has an actual empirical counterpart, but rather to document the kind of situations in which the actual behaviour is close enough to the behavioural insights from

theory (price-taking behaviour in our example) for the model to actually make sense. Section 4.4.2 below provides an example of how experiments can be used to pin down the behavioural content of theoretical assumptions. Last, related to this point, experiments can also be used to assess the robustness of theoretical models in environment variations that are possibly not covered by the model – e.g., socio-personal characteristics of agents, norms or moral values associated with the situation, etc. This serves not only to document the scope of a model, i.e. the extent to which it accurately describes real-world situations, but also to identify parameters or dimensions that are influential on outcomes. When such unexpected influences have been identified in the laboratory, it can stimulate extensions of the theory.

4.4.2 *Case Study*: The Empirical Content of Collusion Theory

Among the most unrealistic, yet often made, assumptions in economics is the idea of infinitely lived agents and the resulting infinitely repeated games. This assumption may change dramatically the predictions of a repeated game. For instance, Illustration 4.2 shows in a simple collusion model that collusion may become a stable equilibrium outcome under this assumption. The stability of such a collusive outcome depends on whether firms are patient enough to refrain from deviating and whether market size is small enough to preserve high enough profits from collusion.

The result is both important and interesting, but relies on assumptions that make no empirical sense, because no economic agent can be reasonably thought of as infinitely lived. Still, the assumption is very useful, because it makes the model of repeated interactions easy to write down and solve. From an empirical point of view, what actually matters is not that much whether or not the assumption is 'realistic' but rather if it helps to describe empirically relevant situations. To that end, the behavioural mechanisms embedded in the assumptions matter more than their real-world counterpart. This question has sparked some debate in the example of infinitely repeated games. According to Martin J. Osborne, for instance, infinitely repeated games 'capture a very realistic feature of life, namely the fact that the existence of a pre-specified finite period may crucially affect people's behaviour (consider the last few months of a presidency or the fact that religions attempt to persuade their believers that there is life after death)' (Osborne and Rubinstein, 1994, p. 136; this is interestingly one of the issues on which the two authors of the book disagree). In terms of behaviour, the main difference between infinitely repeated games and finite horizon ones is thus whether or not current decisions account for what will happen at the last stage of the game. Translated into an empirical question, 'infinitely repeated' means that people do not take the last stage into account, or at least do not take it entirely into account when they decide what their current behaviour should be.

Unlike the formal assumption used in the theoretical model, this behavioural consequence makes a lot of empirical sense. Based on theoretical analysis, it will lead to drastically different outcomes – in a repeated prisoners' dilemma like the one presented in Section 1.3.1, Figure 1.6, for instance, it leads to non-cooperative decisions if people

take the last stage into account in their current decision-making, to cooperative decisions otherwise. The open question thus is, what kind of environment fosters, or undermines, such a driving force of decision-making?

Normann and Wallace (2012) use this idea to provide an empirical test of the range of experimental situations that replicate the assumption of infinitely repeated games. The experiment looks at subjects' cooperative behaviour in a repeated 2*2 prisoners' dilemma game – the same players play together during all repetitions. Four treatments are defined according to different termination rules. In the first treatment (the KNOWN treatment) the fact that the game will last for 22 periods is common knowledge from the very start. In the second treatment (UNKNOWN), there are 28 periods and this is unknown to subjects. In the third treatment (RANDOM-LOW) there is a 1/6 probability that the game will end after 22 periods. In the fourth treatment (RANDOM-HIGH) there is a 5/6 probability that the game will end after 22 periods.[7]

Figure 4.6 shows the number of players who decide to cooperate over the first 22 periods of each treatment. For all termination rules, the initial responses as well as the time trend during the first 12 rounds are very similar. In the KNOWN treatment, cooperation subsequently decreases as a result of an end-game effect – current individual behaviour becomes more and more strongly influenced by the expected outcome at the last stage of the experiment. At the last period of the game, when it is common knowledge that there will be no further repetition of the game, the rate of cooperation is 50% lower than in other treatments.

Beyond this difference at the last stage of the KNOWN treatment, all four treatments generate the same pattern of cooperative behaviour. This has two important implications.

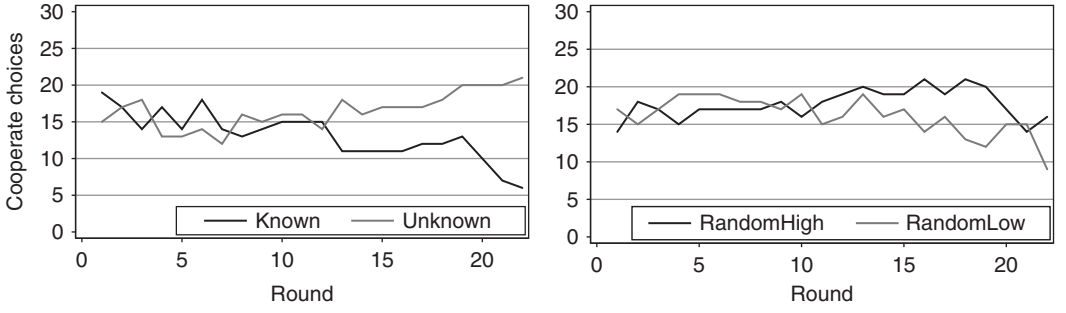


Figure 4.6 Cooperation in repeated games with different termination rules

Note. For each treatment, with varying termination rules, the figure reports the number of subjects who decide to cooperate at each round of a repeated prisoners' dilemma game.
Source: Normann and Wallace (2012, p. 713, Figure 1).

[7] As explained in Illustration 4.2, the random-termination rule used in these last two treatments replicates as closely as possible the model with infinitely lived agents.

First, the choice of a termination rule in experiments on repeated games does not significantly affect individual behaviour. A deterministic, but unknown, number of repetitions, or a random termination rule – or even the first periods of play with a long enough deterministic and common-knowledge termination – all induce subjects to disregard the last stage of the game to the same extent. Second, as shown by the end-game effect observed in the KNOWN treatment, these termination rules all induce the pattern of behaviour expected by infinitely repeated games: higher cooperation supported by the expected rents of cooperating in the future.

4.4.3 The Key Challenge: Internal Validity

Laboratory experiments are well suited to empirically test theoretical predictions because they allow us to build an empirical situation that reduces the environment to only those dimensions that are actually embedded in the model. Experiments thus provide an empirical counterfactual to theoretical models. Testing theory in this context amounts to either comparing behavioural outcomes to theoretical predictions, or performing such a comparison to assess the empirical content of simplifying theoretical assumptions.

In both cases, the process strongly relies on the ability to relate observed behaviour to those features of the environment that aim to replicate the model. If decision-making is rather induced by other dimensions, then observed behaviour has nothing to say about the model itself. In terms of the definitions stated above, this amounts to checking whether or not the experimental outcome $F^m(x^\infty)$ results from the n inputs chosen to replicate the theoretical causal mechanism $f : \mathcal{X}^{n_\mathcal{X}} \to \mathcal{Y}^{n_\mathcal{Y}}$, rather than the $x^{\infty-n}$ inputs at stake in the experiment despite the control. This question is known as the experiment's **internal validity**.

Think, for instance, of an experiment with two treatments, with each session of the first treatment being scheduled early in the morning and all sessions of the other being scheduled in the afternoon. The slots may have an effect on both the kind of subject who shows up to the experiment and the degree of attention and focus during the session. Differences in behaviour between treatments will in this case not only reflect the treatment effects, but also this unwarranted variation between the two environments. In this example, the experiment has serious flaws in terms of internal validity. The main consequence is that the causal inference between the environment and the observed behaviour is challenged. Internal validity refers to how appropriately the causal relationship from inputs to outputs is measured, thanks to the design of the experiment. Since experiments testing theory aim to identify the theoretical causal mechanism, internal validity is the primary challenge.

This definition makes clear that the internal validity is a matter of identification in exactly the same sense as it has been defined and discussed in Chapter 3. When designing an experiment, the aim is to have the subjects' choices induced by the environment chosen, rather than by any other uncontrolled dimensions (such as the subjects' own understanding or interpretation of the game). Internal validity is challenged if the experimenter measures the consequences of a confounding factor rather than a proper causal

effect, rendering the inference based on observed behaviour either invalid or meaningless. The reference to identification helps us understand why the issue of internal validity is hard to tackle. Recall that the quality of identification relies on the exogeneity of the identifying variations, and such exogeneity can never be either proven or empirically tested. In just the same way, it is easy to define what a perfectly internally valid experiment would look like: in such an experiment, all inputs beyond the ones of interest would not interact with subject's responses to the controlled inputs – hence being exogenous. But it is far less easy, and in fact impossible, to definitely prove that an experiment is internally valid. We will devote a dedicated chapter to this issue (Chapter 5), and discuss how experiments can be designed and implemented in a way that enhances their internal validity.

4.5 How Experiments Interact with Reality: Searching for Facts

Experiments searching for facts seek to produce empirical knowledge on situations for which theory has little or nothing to say: either because no theory exists or because the existing theory makes predictions that are obviously inconsistent with behaviour. In these cases, experiments can be used to 'establish and document stylised facts, in the form of either observed phenomena or observed causal effects' (Schram, 2005, p. 232). The ability of experiments to test-bed such facts stands on the other side of the continuum between theory and reality. Experiments allow us to build pseudo-real situations, of which the set of inputs can be chosen in accordance with the main features of the situation under study. The experiment then provides an empirical understanding of the kind of behaviour, decision and outcome induced by a given environment. Auction mechanisms, and the kind of bidding behaviour they induce, can, for instance, quickly become highly technical from a theoretical point of view. Once the boundaries have been reached of the ability of theory to actually predict how bidders will behave when faced, e.g., with a given set of rules, and under a specific information set, experiments searching for facts can produce useful knowledge. To that end, it is enough to experimentally build the auction situation and observe both the bidding behaviour and the market conditions it gives rise to. Illustration 4.5 describes a typical example of a well-known experiment searching for facts.

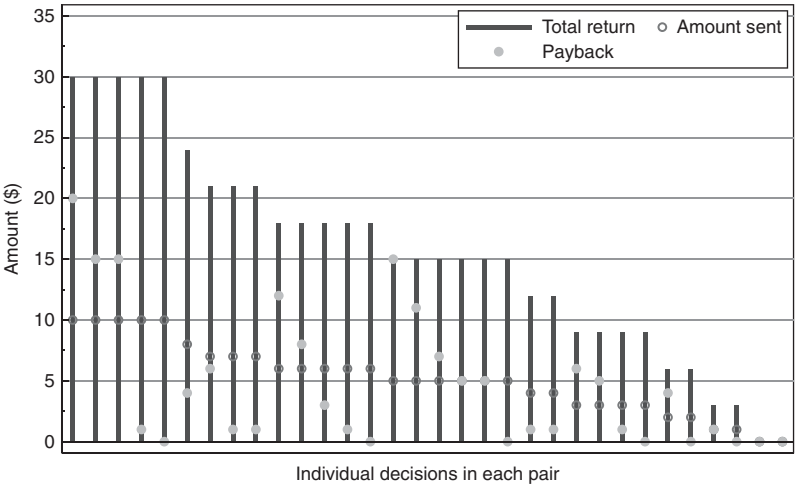
An important feature of experiments serving this purpose is the ability to replicate the results. The more often a given behaviour arises in subsequent implementations of a given experiment (possibly with slight variations), the more these observations become actual regularities, giving rise to stylised fact. This is one reason for using meta-analysis (see e.g. the results of the dictator game presented in Section 4.3 for an example, and Chapter 8, Section 8.4, for a detailed discussion).

4.5.1 The Key Challenge: External Validity

Accumulated evidence from such experiments aims to produce empirically based knowledge about the behaviour and outcomes generated by a given environment. Internal validity is obviously a necessary condition for this empirical knowledge to be

Illustration 4.5
Trust: evidence from the lab

The trust game has been purposefully introduced by Berg et al. (1995) to provide empirical facts on the existence and extent of trust and trustworthiness in economic relationships. In this experiment, each subject first receives a $10 windfall endowment. Subjects are then randomly split into two groups, defining their role as either senders or receivers. Senders have to decide to give any share (including 0) of their $10 endowment to an anonymous, and randomly matched, receiver. The experimenter then triples any amount sent by the sender to the receiver – i.e. the actual amount a receiver receives is three times the amount sent by the sender. Last, receivers are asked which part of the tripled amount they want to send back to the sender. Standard economic theory, based on self-regarding preferences, allows for neither trustworthiness nor trustfulness. Applied to this situation, it thus predicts that no cooperation should occur: the sender should keep everything, because the receiver is expected to return nothing. But this game is built in such a way that the rent from trust is huge: the question is thus how much of this economic value, created by trust, can be achieved by human beings despite the incentives to behave selfishly. The answer from simple economic theory is way too extreme to be informative. The results from the experiment, presented in the figure below (from Berg et al., 1995, p. 130, Figure 2), confirm that a large share of this benefit is actually realised.



The figure shows the decisions made in each of the 32 pairs of subjects, in decreasing order of the amount sent (white circles) and the resulting amount received (height of the bars). The amounts sent back are shown with black circles. On average, senders 'invest' about 50% of their endowment in a transfer to the receiver. There is a large heterogeneity in terms of what receivers sent back. About 20% of receivers send back no money at all, while a large majority send back something. Trustworthiness thus seems less widespread than trustfulness. The main outcome from these two behaviours is that the return to trust is on average 0: around 95% of what is invested (from senders to receivers) is repaid. At the same time, the average total return is $15 – from an endowment equal to 10. Trust is thus beneficial for the economy as a whole, leading to a 50% increase in the monetary value to be split between players.

sound – as the experiment would otherwise document the effect of inputs different from the ones under study. But, in the case of experiments searching for facts, it is far from being the end of the story. If the experimental outcome is actually conclusive, then it provides observations on the behaviour of those individuals who participated in the experiment, facing the artificial institutions built in the laboratory. Each element of the sentence can be a matter of concern. Do the experimental subjects behave in the same way as the actual economic agents would? Isn't the game too abstract to induce the same kind of behaviour that would occur in the real world?

Answering this kind of question is of utmost importance in the case of experiments searching for facts – just because these facts are meaningless if they have nothing to do with real-world situations. This issue is known as **external validity**, in reference to what happens outside the experimental environment. The two kinds of validity refer to what can be made of the laboratory observations, but while internal validity refers to the quality of the empirical measure generated by the experiment, external validity is rather related to its relevance. The strength and range of external validity of the results of an experiment thus mainly condition its interpretation. The question is no easier to deal with than that of internal validity. But the reason why it is the case is quite different. The answer to the external-validity question is in a sense as simple as whether or not it is the case that the causal relationship measured in the experiment is specific to the laboratory context or would occur as well in the real world. What makes the question hard is the many ways in which the words '*specific*', '*as well'* and '*real world*' can be understood when assessing external validity. This discussion, and what is currently known about the external validity of laboratory experiments, will be the topic of Chapter 8, of which the case study below provides an example.

4.5.2 *Case Study*: Testing the Reciprocity Model in the Field

The empirical evidence from the trust game (presented in Illustration 4.5) echoes the Fehr et al. (1993) experiment in support of the fair-wage-effort hypothesis. Both potentially have huge consequences for the understanding of labour contracts. From these results, it is no longer true that flat wages are unable to foster performance at work. It strongly widens the set of incentive-compatible compensation devices. This is indeed one of the reasons why this behaviour has been so widely studied in the economics literature. But this all is true only if it is actually the case that this kind of behaviour occurs in actual work relationships – if these results have external validity.

Gneezy and List (2006) offer an empirical investigation of this question – 'is the behaviour of laboratory subjects, who are asked to choose an effort or wage level (by circling or jotting down a number) in response to pecuniary incentive structures, a good indicator of actual behaviour in labour markets?' (p. 1366). To that end, one needs to define what a labour market is in the real world, i.e. what makes it specific as regards the behaviour studied in the laboratory. Gneezy and List focus on the duration of the work relationships.

The empirical investigation relies on an experiment implemented in the field. Students are recruited through advertisements to computerise the holdings of the university's library. The advert announcing the experiment offers a $12 wage per hour of work, so that this wage rate is known by all the students who come to participate. In the control group (NO GIFT treatment), students who come on the morning of the experiment are invited to sit in front of a computer and are paid $12 per hour. They work for a total of six hours, and the number of books correctly entered into the system is recorded by the computer. This number is an observable measure of work performance and is used as the main outcome variable of the experiment. In the treatment group (GIFT), the experiment works exactly the same except for one feature. Upon arrival, students are told that the wage rate has been revised upward to $20 per hour. In terms of Akerlof's gift-exchange model, the initially announced wage sets the reference wage of subjects coming to participate. The good surprise implemented in the treatment thus replicates the fair-wage condition of the model.

The main results observed in the experiment are shown in Figure 4.7. The lines are drawn separately for each group according to the duration of the experiment and show the evolution of the performance measured every 90 minutes. In the first 90 minutes, the treated subjects in the treatment group produce around 25% more output per hour than those in the control group, which is consistent with the standard results on the fair-wage-effort hypothesis. In the next portion of 90 minutes, the difference falls to 10%, and becomes (almost exactly) 0 afterwards, in such a way that the performance of 'workers' is now the same whatever the level of the fixed wage they are offered. The experiment thus shows that the effect of a higher fixed wage eventually fades away as the duration of the contract increases. This is obviously a strong limitation to the ability of the fair-wage-effort hypothesis to describe work relationships, and is actually taken by Gneezy and List as strong evidence against the external validity of experiments supporting the existence of such behaviour.

In response to Gneezy and List, Falk (2007) notes that, strictly speaking, the external validity of experimental results is satisfied if the consequences are the same when the inputs controlled for in the experiment are also at stake in the field. To make the point, Falk offers an alternative test of the external validity of these results, based on a one-shot interaction. The experiment is designed jointly with a charitable organisation, whose aim is to help children in need. The experiment consists of sending solicitation letters to a random sample of households in Zurich (Switzerland). The letters ask for donations for funding schools for street children in Dhaka, Bangladesh. The households are randomly assigned to three treatment groups. The first group receives only the solicitation letters asking for donations. The second group receives the letter and a 'small gift' – a nice postcard. The third group gets the letter and a pen, which represents a 'large gift'. The letter makes it clear that the presents are free and for the recipients to keep, regardless of whether they decide to donate or not. The experiment thus replicates a one-shot real-world fair-wage-effort relationship: the solicitation letter stands for a work contract; the amount of the donation is a non-contractable effort; and the gift, when there is one, is an unexpected compensation for this effort.

Table 4.1 Gift exchange in the field: donation patterns

	No gift	Small gift	Large gift
Number of solicitation letters	3.262	3.237	3.347
Number of donations	397	465	691
Relative frequency of donations	0.12	0.14	0.21

Note. For each treatment group in a column, the table reports the sample size and the the number of households who donate in return to the letter.
Source: Falk (2007, p. 1505, Table 1).

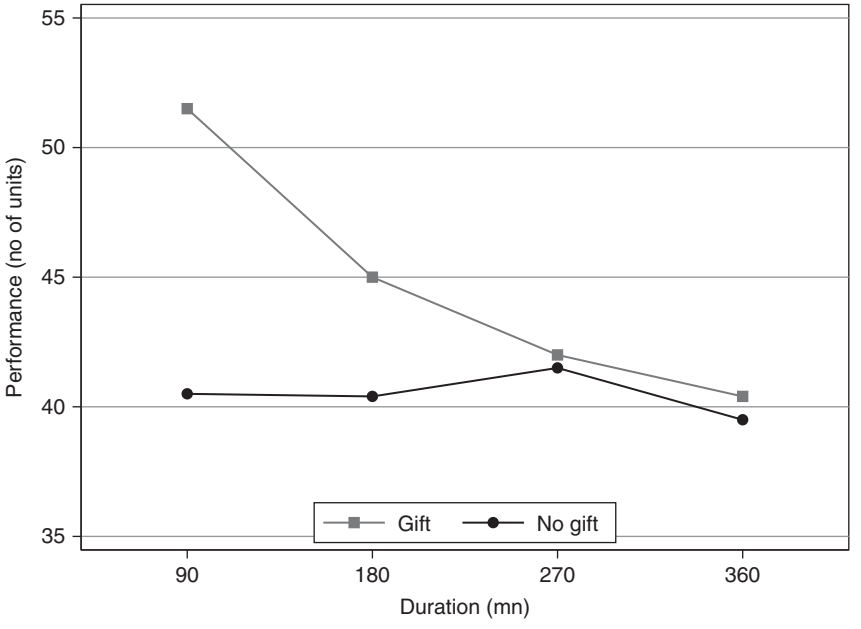


Figure 4.7 Reciprocity in the field

Note. The figure shows the average performance (number of books entered in the system) measured every 90 minutes, respectively in the control (NO GIFT) and in the treatment group (GIFT) according to the duration of the experiment.
Source: Gneezy and List (2006, p. 1371, Figure 1).

Table 4.1 reports the donation patterns observed in each of the three treatment groups. Clearly, donation frequencies increase with the inclusion of a gift as well as with the value of the gift. While the donation frequency only slightly increases when a postcard is associated with the letter, it almost doubles when the letter includes a bigger gift. This increase in the extensive margin of the donation does not crowd out the amount of the donations: no treatment effect shows up when comparing the distribution of the amount donated in each treatment (see Falk, 2007, Figure 2, p. 1506).

Because they happen in the field, and involve real-word decisions in a real context, these results substantiate that laboratory behaviour has some external validity. The same kind of behaviour observed in the laboratory is generated by the same set of outputs.

This is actually consistent with the Gneezy and List (2006) results, if one focuses on the first 90 minutes of the experiment – the maximum duration of the target task of the vast majority of laboratory experiments. Both these results thus confirm that the gift-exchange mechanism is one driving force of real-word economic behaviour, although it is not the case that the mechanism works under all possible circumstances – it is strongly sensitive, in particular, to the duration of the relationship.

Beyond the case of the gift-exchange mechanism itself, these results illustrate how controversial the question of external validity can be. The good news, however, is that in the end it always boils down to an empirical question. It thus stimulates an informed debate involving laboratory studies to assess the robustness of observed behaviour to alternative inputs in a highly controlled environment as well as field experiments to confront such effects in real-world behaviour. The gift-exchange model is a typical example of how the experimental literature evolves over the years according to this process. The experimental research on reciprocity has originated from theoretical models (Akerlof, 1982; Akerlof and Yellen, 1990; Shapiro and Stiglitz, 1984), which were then put to the test in laboratory experiments. These results were then challenged and tested for robustness through field experiments. This process is still ongoing.

Summary

The core of economic science is to understand social reality based on theoretical models. Experiments are central to this process, contributing to each of its directions. Following Roth's (1988) seminal classification, experiments can serve three different purposes: testing theory, i.e. assessing the empirical relevance of theoretical models; searching for facts, by documenting situations that are ill-covered by economic theory; or supporting the design of public policies, which is a combination of the first two. The ability of experiments to achieve these goals raises the question of the interaction between theory, experiments and reality and how they inform one another. This chapter introduced an integrated framework on what an experiment, a theoretical model and reality are, showing that both theory and experiments are restricted environments designed to simplify reality – a must-have to be able to understand it despite its complexity.

This is the building block of a discussion of how theory and experiments, together or separately, inform our understanding of the real world. First, testing theory in the laboratory amounts to building an empirical counterfactual to the theoretical causal mechanism. This is achieved based on causal inference between observed behaviour and the institutions purposefully implemented in the laboratory in order to replicate the model's assumptions. As a result, the big challenge faced by these kinds of experiment is their internal validity, i.e. whether or not observed behaviour is induced by the chosen institutions rather than by uncontrolled dimensions. This issue is at the heart of how experiments are designed and put in practice. This is the topic of Part III. Chapter 5, in particular, describes both the main impediments to internal validity and how to solve them. The practicalities of experiments, described in Chapters 6 and 7, put these principles in practice.

Internal validity is obviously important as well for experiments searching for facts. They aim to provide stylised facts about situations that are poorly covered by economic theory. This amounts to creating a pseudo-real situation, focusing on a few dimensions of interest of the environment, in order to document the outcomes and behaviour they generate. They can be seen as an empirical model, with observed behaviour standing for the predictions. But such experiments are informative about real-world mechanisms only if what happens in the laboratory also happens outside – i.e. if the experiment is externally valid. This raises the question of what experiments tell us, the focus of Part IV. It opens with a focus on the real world in Chapter 8, addressing the question of the external validity of the results raised by laboratory experiments. The will to challenge and refine external validity stimulates an empirical process going back and forth from the laboratory to the field in order to stabilise and refine empirical knowledge. The final stage of this process is eventually to close the loop and get back to theory, to adjust for the empirical phenomenon it pinpointed. How such induction can and may occur is discussed in the first part of Chapter 9. Then, ultimately, the experimental empirical phenomenon becomes part of the toolbox of the economic analysis of the outcomes generated by different kinds of institution. This serves as a basis for policy design. The second part of Chapter 9 focuses on this third aim of laboratory experiments, discussing how well-designed and externally valid experiments, either testing theory or searching for facts, improve our understanding of public policies.

Part III

How? Laboratory Experiments in Practice

5 Designing an Experiment: Internal-Validity Issues

Discussing the need for experiments in the previous part delivered two take-home messages. First, from an empirical point of view, an experiment allows us to choose the data-generating process – the properties of which are the core of the inference properties of any empirical strategy. Second, at the same time, an experiment is also a pseudo-real situation which shares features with both theoretical models – some of the driving forces of behaviour are chosen – and the real world – there will always be some feature that remains beyond control but nowadays influences behaviour. The aim of this chapter is to operationalise these two observations by describing how the DGP can be chosen in such a way that identification is achieved despite the inevitable uncontrolled driving forces of behaviour.

This concern is often referred to as the internal validity of the experiment: does the experimental environment produce convincing outcome measures? Answering this question comes down to asking how experimenters can create a world in the laboratory that best fits their observational needs. As a result, this chapter will also be about how to make the laboratory best suited to its measurement objectives: how do we concretely **design** an experiment? What are the main concerns and pitfalls? What are the choices and trade-offs to be made? To facilitate the discussion, it will be helpful to more precisely describe the components of the experimental DGP, which is the aim of Section 5.1. This will help understand more precisely how the internal-validity issue arises, and how it can be dealt with. This will lead to two complementary answers: internal validity requires controlled dimensions to drive outcome behaviour, and uncontrolled ones not to be confounding. The main feature of experimental designs used to fulfil the first dimension is the use of monetary incentives, which we describe in Section 5.2, and the implementation of exogenous changes through experimental treatments (Section 5.3). We then move to the features that are likely to induce uncontrolled and confounding variations: the perceived experiment induced by how the experiment is described to human beings asked to behave in the experiment, and beliefs about others' behaviour. Through this review, this chapter will thus describe the most crucial best practices in the implementation of experiments, and discuss their rationale.

5.1 What Is an Experiment? How Is It Linked to Internal Validity?

In one of the classics of the methodological literature in experimental economics, Smith (1982) defines an experiment as a 'microeconomic system' made up of three

components (the environment, the institutions and the resulting behaviour). While the terminology is different, this definition is very close to the one we introduced in Chapter 4 (Section 4.2.3). The first two components – the environment and the institutions – are nothing but a partition of what we introduced as *inputs*, deciding on the pseudo-real situation that subjects are faced with. The dividing line between these two components is the following: the environment encompasses all the initial circumstances of the experimental system, while the institutions frame its dynamic evolution.

The last component is the same as the transformation function, through which these inputs result in specific occurrences of the experimental measures. Each definition serves its own purpose. In Chapter 4, the definition helped contrast experiments with the two main objects of economic science – real economic life and theoretical models. The main contribution of Smith's definition, in terms of components, is to describe more precisely the ingredients involved in the choice of inputs, and how influential they are on laboratory outcomes. It emphasises that an experiment is a closed system. This will prove very helpful to discussing internal validity – the accuracy of the link between the chosen structure of an experiment and the decisions it elicits from subjects, and how to best choose this structure in that regard.

5.1.1 Experiments as 'Microeconomic Systems': The Components of an Experiment

The *environment* is the collection of all characteristics describing what the system is made of. Important pieces in this collection are: the number of agents (players in a game, buyers and sellers on a market, etc.), the specification of the commodities (tokens and their face value in a trust game, abstract good in an induced-value auction, etc.), and agent-specific endowments in terms of resources, preferences (over allocations, i.e. utility functions) and technology (e.g. skills and knowledge). This defines the givens of the system, some of which are individual-specific. As such, they might be private information.

The second constituent of a microeconomic system is made of the *institutions*, which define the functioning of the system. This first amounts to specifying the ways agents act together: how they communicate and decide (what is the set of available messages, what is the order in which they are decided) and how they interact (who knows what, and when) within the environment. The consequences of these actions for the state of the system are driven by the allocation rules set by the institution. This determines how the initial endowments are affected by agents, messages and decisions, and how property rights over this allocation are distributed among agents. This is coupled with cost-imputation rules, specifying how agent resources are impacted by the change in the allocation. Last, the dynamic of the system is decided by the set of adjustment-process rules, including the initial rules (how the system is initiated), the transition rule (how messages drive the system from one state to another) and the stopping rule (deciding when the exchange of messages is terminated).

These elements are general and precise enough to characterise any microeconomic system. Applied to a laboratory experiment, they highlight the complete set of characteristics that are to be decided on when 'designing' or 'building' an experiment.

Focus 5.1

Cold versus hot: available measures of outcome behaviour

One important dimension on which controlled experiments enhance the observation possibilities is the set of decisions elicited from subjects. In sequential games, the most natural way of eliciting decisions in a game is to ask subjects to make a choice when they have to. In a seminal paper, Selten (1967) introduced an alternative elicitation scheme called the *strategy method*. It amounts to asking subjects to post the full set of contingent actions they would make at any possible node in the game. The outcomes are then determined by having each subject's full set of actions play against one another. Consider, for instance, the four-moves centipede game presented in Chapter 1, Section 1.3.2. Applying the strategy method would amount to asking Player 1 whether Take or Pass would be chosen at nodes 1 and 3, were it reached in the course of the game; and similarly Player 2 about nodes 2 and 4 – both without knowing anything about the choice of the other. The actual outcome for these players then results from the intersection of their contingent plan of actions. The strategy method thus widens the scope of observed outcomes to choices that never have to be actually made. The two methods do not exactly coincide in terms of the driving forces of behaviour they elicit: direct answers are 'hot' – decided spontaneously as the decision problem arises – while decisions elicited through the strategy method can be seen as 'cold' – they force subjects to consider all possibilities at once (Brandts and Charness, 2000). The two methods can easily be compared within an experiment: it amounts to having different subjects play the exact same game, but under each of the two elicitation methods. Brandts and Charness (2011) provide a literature review of existing comparisons and show that little quantitative difference, and no qualitative variation, are generally observed. When a difference is to be expected, a choice needs to be made about the accuracy of either of the two methods to best answer the empirical research question.

The ability to choose the specification of each and every of these components is what makes experiments a highly controlled empirical setting: deciding on the environment and the institutions amounts to deciding on the specification of the microeconomic system. But this same control over the system is also what makes the empirical evidence highly sensitive to the accuracy of this choice. As in any microeconomic system, an experiment is closed by *agents' behaviour*. Under the rules set by the institutions, and the endowments set by the environment, the state of the system evolves according to agents' individual decisions. This includes two different kinds of outcome: the final state of the system, reached thanks to all previous decisions and interactions; and agents' response behaviour, governing individual reactions in the course of the experiment. Both are the behaviours elicited by the system, which can thus be seen as the empirical reaction functions of the experiment's subjects to the environment and the institutions they faced in the laboratory. As explained in Focus 5.1, several methods are available to design the measures of outcome behaviour.

This view of an experiment as a microeconomic system helps characterise the empirical approach. The experimenter has control over the environment and the institutions, which together result in agents' behaviour. The aim is to infer the empirical properties of the chosen environment and the institutions from observed behaviour. Such inference

is accurate only if behaviour is actually induced by the chosen microeconomic system. Here stands the core of the internal validity of an experiment. There are two necessary conditions for the experiment to achieve this goal. First, internal validity requires that decisions from subjects occur within the system, i.e. that behaviour responds to the chosen microeconomic system. The remainder of this chapter describes the building blocks of how experiments are designed to that purpose. Second, this is not enough to achieve proper inference if uncontrolled dimensions occur in a way that is confounding. The design choices of experiments aim to fulfil both conditions.

5.1.2 Internal Validity and the Design of Experiments

Because it is a matter of inference, internal validity shares a lot with the idea of endogeneity in econometrics, as introduced in Chapter 3. Identification is challenged as soon as unobserved variations contributing to the outcome occur at the same time as experimental controls. In the framework of the estimation of causal treatment effects, for instance, the data do not deliver identification of the causal parameter if unobservables systematically change in line with the implementation of the treatment. One take-home lesson from this discussion is that identification is achieved if the true data-generating process – what makes outcomes what they are – complies with some assumptions on the determinants of the outcome variable – what we called the assumed DGP. In an experiment, the true DGP is chosen on purpose, by choosing the microeconomic system described above. In the experimental economics literature, such a choice of the specifications of the system (how decisions are elicited and taken in the laboratory) is referred to as the *design of the experiment*. Proper identification in this context implies choosing the specification of the microeconomic system, the experimental design, in order to comply with identifying assumptions. This is the crucial criterion governing the experimental design: choosing the experimental data-generating process in order to achieve, for the best, proper identification of the relevant parameter(s).

Although it may be disturbing at first glance, 'for the best' in the previous sentence will come as no surprise to the reader aware of the discussion in Chapter 3. Identifying assumptions hold on the mechanisms that actually generate outcome behaviour. There would be no need for empirical research if such mechanisms were either perfectly known or observable. Any effort to achieve identification thus relies on a pre-existing knowledge or understanding of what these actual mechanisms are. In the same way as identification properties of estimators are conditional on non-testable identifying assumptions, internal validity relies on assumed properties of agents' responses to the microeconomic system they face. This remark has two important consequences. First, internal validity always reduces to a matter of faith. One will never be able to prove that the unobservable true DGP actually matches the assumed one (again, in the same way as exogeneity cannot be tested or proven). The reverse is not true, however, as it is enough to show that a confounding effect does have an influence on behaviour to establish that internal validity is challenged. Illustration 5.2 provides an example of such empirical test of internal validity, applied to the WTA/WTP discrepancy discussed in Illustration 5.1 and Focus 5.2. But there is an endless list of such unobservables, so that testing them

Illustration 5.1

Endowment effects in market behaviour

The Coase 'theorem' (Coase, 1960) states that in the absence of any transaction cost, the allocation of property rights does not matter to the efficiency of the final allocation. This is one of the building blocks of public economics. The seminal experiment by Kahneman et al. (1990) provides strong evidence against this principle. In this market experiment, the subject pool is divided at random into two sub-populations: subjects are sellers in the first one, buyers in the second. We herein focus on the last four periods of the experiment, in which sellers receive a coffee mug. The market value of this object is $6 in the university book store at the time of the experiment. Sellers are asked to state the minimum price they need to receive to agree to sell the good – their *willingness to accept* (WTA) – while buyers are asked to state the maximum price they would like to pay to acquire the good – their *willingness to pay* (WTP). Based on elicited answers, the market is cleared and transactions are accordingly implemented. The main results from this experiment are summarised in the table below (Kahneman et al. 1990, p. 1332, Table 2) displaying the average price chosen by buyers and sellers.

Trial	Trades	Price	Median buyer reservation price	Median seller reservation price
	Mugs (expected trades = 9.5)			
4	3	3.75	1.75	4.75
5	3	3.25	2.25	4.75
6	2	3.25	2.25	4.75
7	2	3.25	2.25	4.25

Since the good is the same and subjects are allocated randomly to groups, prices should – by design – be the same in both groups in a Coasian world. This is by far not the case: the WTA is more than twice the WTP. The behavioural interpretation of this discrepancy is known as the endowment effect: the property of the object generates value on its own. This contradicts the Coase theorem, as the initial allocation of property rights matters for the final allocation through market transactions.

all one after the other is a hopeless avenue. Second, for this same reason, an experiment will never be 'perfectly internally valid'. The best one can do is to choose the design as carefully as possible so as to (i) make it likely that the chosen environment does matter for behaviour and (ii) discard influences that are likely to be confounding.

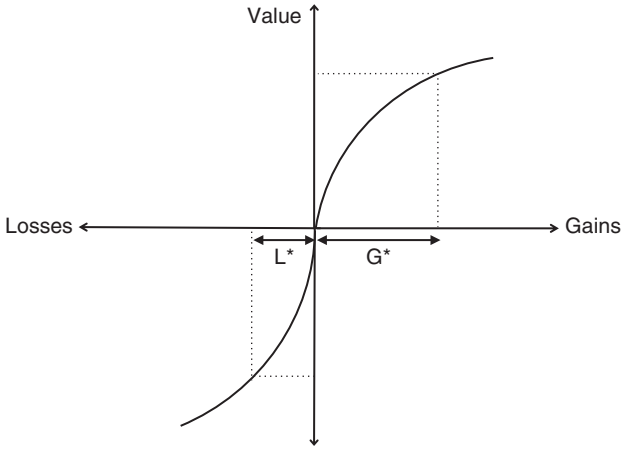
5.1.3 Indirect Controls: Block Everything You Can, Randomise Otherwise

Despite the wide scope of controls offered by the experimental environment, there always remain many features of the context of the decisions that are uncontrollable. This is highlighted in Chapter 4's definition of what an experiment is: no matter how large the set of the controlled inputs, the actual input of an experimental situation is

Focus 5.2

Loss aversion: a behavioural foundation for the endowment effect

The endowment effect giving rise to the WTP/WTA discrepancy shown in Illustration 5.1 can be rationalised by a model of loss aversion (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), according to which subjects dislike losing what they already have even if traded against something else of the exact same value. For example, Knetsch (1989) elicits the willingness to exchange goods (the experiment is based on exchanges between candy and a mug) and shows it strongly depends on initial endowments. Indeed, 89% of those initially given a mug opted to keep it while only 10% of those initially given candy opted to exchange it. As a consequence, exchange appears as a loss of the endowment and is rejected by most of the subjects, whatever their initial endowment. For example, in Kahneman et al. (1990) the value of a good is much higher for sellers than for buyers, because the former ask to be largely compensated for their loss. When an individual is loss-averse, 'losses loom larger than gains': losses are weighted much more heavily than objectively identical gains in the evaluation of prospects. This occurs whether such prospects are risky or not. Loss aversion results in a utility function that is steeper for losses than for gains. A common graphical representation of loss aversion is shown in the figure below. When facing a loss x_L, a much larger gain $x_G >> x_L$ is needed to compensate the individual from the negative value associated with the loss.



This utility function is reference-dependent in the sense that gains and losses are defined relative to a reference point. In the figure, the reference point is set to 0 and the utility function (which is sometimes called the 'value function' in reference-dependence models) is assumed to be concave for gains and convex for losses. This shape corresponds to a diminishing sensitivity towards gains and losses. The kink at the reference point represents loss aversion: as the individual valuations depend on a reference situation, the asymmetric shape induces loss aversion and therefore an endowment effect. Subjects prefer to keep what they have than to lose it in favour of something of the same value.

Illustration 5.2

Identified failures of internal validity: misconceptions about the endowment effect

In a series of articles, Plott and Zeiler (2005, 2007) question the internal validity of the endowment effect observed in market experiments. Based on a literature review of existing experiments, four main dimensions of the experimental design are found to influence the WTA–WTP gap:

- the elicitation device (in particular, whether or not it is incentive-compatible),
- the extent of subject's training with the mechanism,
- whether practice rounds are paid, and
- whether anonymity is ensured.

These features are hypothesised to affect subjects' understanding about the environment. To investigate whether misconceptions might occur when some of these features are missing, the experimental design consider them all at the same time in a replication of the Kahneman et al. (1990) mug experiment described in Illustration 5.1. The table below displays the main result from the study (from Plott and Zeiler, 2005, p. 539, Experiment 1, Table 4).

	N	Individual decisions (in USD)	Mean	Median	Std. dev.
WTP	15	0, 1, 1.62, 3.5, 4, 4, 4.17, 5, 6, 6, 6.5, 8, 8.75, 9.5, 10	5.20	5.00	3.04
WTA	16	0, 0.01, 3, 3.75, 3.75, 3.75, 5, 5, 5, 6, 6, 6, 7, 11, 12, 13.75	5.69	5.00	3.83

This implementation of the experiment gets rid of any evidence of a WTA–WTP gap: in terms of mean price, median price and dispersion, all outcomes are very similar when elicited from either buyers or sellers – all comparisons are non-significantly different. This result is highly controversial given the influence of the endowment effect on both the theoretical and empirical literature. The nature of the good, in particular, has been shown to be crucial as the gap seems more robust to the procedures when lotteries, rather than consumption goods, are exchanged on the market (e.g. Isoni et al., 2011; Plott and Zeiler, 2011).

$x^{\infty} = x^{n} \cup x^{\infty - n}$, because it belongs to the real world (see Section 4.2.3). Leading examples of such inputs are subject-specific heterogeneity, like their beliefs (about the behaviour of others but also, for instance, about how trustable is the information coming from the experimenter), or their mood or emotions when arriving; but this also includes experimenter-specific heterogeneity (e.g. how clear is the reading of the instructions, how 'serious' or trustable the experimenter seems to be), lab-specific characteristics (location, comfort, etc.). All these examples are sources of noise in the experimental observations: decisions will likely not be the same, within the same experimental design, when either of these features changes. They are also both unobservable and generally impossible to control – one cannot choose to implement a given level of trust towards the experimenter: subjects are endowed with their own, which can hardly be measured. But such noise is not confounding per se. As such, it only affects statistical inference

(the precision of the estimated effects). Reducing the intensity of the noise improves the quality of experimental outcomes, by delivering more precise estimates.

But what matters for identification is whether such noise is correlated with the variables of interest. To make things concrete, take an experiment that aims to measure gender effects by comparing behaviour between only-males and only-females versions of the experiment. If the male version always take place before lunch, and the female version just after, this very fact might induce noise in observed behaviour. But the chosen implementation moreover generates a correlation between noise and gender. The observed differences in outcomes between the two versions will not be an accurate measure of gender effects. It is so because changes in the outcome (through the noise) occur at the same time as variations of the target variables, hence misleadingly suggesting a relationship between the two while both variations are in fact caused by the noise. Such a correlation between the noise and the outcome is said to be confounding and challenges identification. The concern for internal validity leads to preventing correlations between the noise arising in the experiment and the variables of interest. While there are as many internal-validity issues as the number of known or expected confounding mechanisms, two kinds of strategy introduced in Section 3.4 circumvent them. The best practices in the design of experiments operationalise these strategies.

'*Blocking*' strategies aim to hold constant nuisance dimensions of the experiment: nuisance is still there, and is not observed in any way, but since it no longer varies, it is no longer confounding – hence implementing the identification strategy described in Section 3.4.3. Blocking amounts, for instance, to avoiding using several different physical laboratories to run several sessions of the same experiment, or changing the identity of the experimenter. The more such features remain the same, the more likely it is that changes in outcome behaviour are immune to their effect, because they hold constant across all instances in which the outcome variable is observed. For this same reason, this also achieves better precision in the measure of the relationship of interest.

The alternative design strategy, '*randomisation*', is used when variation in nuisances cannot be avoided. As shown in Section 3.4.1, if such variations happen but are uncorrelated with the target determinants of behaviour, they induce noise in the data and less precision in the econometric analysis, but they are no longer confounding. This concretely implies choosing the value taken by these nuisance variables according to a random draw. Many dimensions of an experiment can be chosen in this way (and the general principle is to follow a random-allocation rule in all instances in which blocking is not available). This is the reason why, in particular, computers in the laboratory are allocated to subjects by asking them to draw an assignment card before entering the lab – in such a way that they do not choose where they sit, and who the neighbours are. Similarly, in those experiments featuring different kinds of position in the game (like sender/receiver, for instance) these roles are not attributed based e.g. on the location of the computer in the room, or to subjects based on their arrival order, but rather by random assignment across all subjects to the session – in such a way that any systematic relationships between location-specific or subject-specific heterogeneity and role in the game are broken.

On a final note, it is worth stressing that any choice of an experimental design can only be assessed as regards the specific research question the experiment aims to address. The research question is what defines the outcome, the noise, and relevant variables of the experiment – hence what might, or might not, be confounding. As an example, when experiments seek to measure the causal effect of changes in the environment or institution, the outcome variable is the difference in behaviour between two settings. The relevant noise is those unobservables that make the *difference* in outcome change at the same time as the relevant *change* in the context. Basically, any noise influencing the outcome levels in the same way in the two settings cannot be confounding. In contrast, the case study below provides an example of a *measurement experiment*, in which observed behaviour per se is the outcome of interest. In a measurement experiment the identifying assumptions are more demanding, because any unobservable influencing behaviour belongs to the relevant noise term, and might thus be confounding. As suggested by these examples, a sensible choice of design for one experimental investigation can thus be just obviously wrong for another one. Consequently, the insights developed in this chapter should be seen as neither absolute principles any experiment must comply with – as a matter of fact, the chapter will describe plenty of counterexamples to the general discussion – or an exhaustive review of confounding effects found in experiments – as each new design is likely to raise its own. They describe the set of tools available to undermine the effect of usual suspects challenging internal validity, and illustrate the practical consequences of the principles discussed here on how experiments are designed.

5.1.4 *Case Study*: A Measurement Experiment: The Voluntary-Contribution Mechanism

Public goods in microeconomic theory share two specific features: they are non-rival and non-excludable. Non-rivalry means that any unit of the good that is consumed by an economic agent still remains available for consumption for another one; non-excludability happens when there is no way to prevent an agent consuming available units of the good. National security is a prototypical example of a public good. First, it is non-rival because one person enjoying domestic safety does not hinder another person from 'consuming' this same safety. The important consequence in terms of microeconomic analysis is that the cost of providing national security to an additional inhabitant is 0. This stands in sharp contrast with private goods, for which serving more consumers requires producing more of the good. Second, national security is also non-excludable, as anyone living in the area will benefit from it. Again, this is a huge departure from standard analysis of private goods, for which consumption can be made conditional on paying a price for each unit of the good that is consumed.

The main consequence of these two features (together, as none of them alone is enough to define a public good) is that the allocation of public goods is a typical example of a market failure: the number of units produced and consumed in the economy if economic agents behave in an unconstrained and decentralised way is not the best they can achieve together with the available resources. The intuitive reason for that is easy to understand: because of non-rivalry, the number of units that should be produced is determined by the sum of all consumers' willingness to pay for each of these units – because

all consumers will then be able to consume each and every unit of the good. To achieve such a level of production, each consumer should thus be asked to pay an individualised price exactly equal to one's own preferences towards the good. But because of non-excludability, consumers can enjoy any unit of the good once available in the economy at no cost – just because there is no way to constrain people to pay for a non-excludable good. On a free market, everybody will thus hope to rely on others to pay for the production cost of the good, while enjoying those units that eventually become available. This obviously results in no production at all. In behavioural terms, the key mechanism in this reasoning is free-riding behaviour: if asked to freely choose whether or not they want to contribute to funding the production of a public good, rational consumers will give the least possible. The empirical content of the microeconomics of public goods, and the design of institutions aimed to enhance the allocation, crucially depend on the relevance and extent of such behaviour.

The Voluntary-Contribution Mechanism (VCM) is an experimental game purposefully designed to provide an empirical measure of free-riding behaviour (Isaac et al., 1984). This game gathers N players who each receive an endowment denoted e_i. Each player is asked to decide on the allocation of this endowment between two possible investments: a private or a public one. The per-unit individual return of the private good is q (> 1): each dollar put by individual i in the private investment increases i's earning by $\$q$. The public good, by contrast, benefits anyone: the return on each dollar invested in the public good is Q but this amount is equally split between all members of the group, increasing the individual earnings of each one of them by $\$Q/n$. We denote ℓ_i, with $0 \leq \ell_i \leq e_i$, the level of 'contribution' to the public good (the amount allocated to the public investment). Once all allocation decisions have been made in the group, the individual payoff resulting from them is thus:

$$\underbrace{q(e_i - \ell_i)}_{\text{Return from private good}} + \underbrace{\frac{Q\left(\ell_i + \sum_{j \neq i} \ell_j\right)}{N}}_{\text{Return from public good}}$$

As simple as it is, this game replicates the main features of the social dilemma raised by public-good provision in an economy. It is usual to refer to $Q/(qN)$ as the MPCR, marginal per capita return. As soon as the returns and group size are such that $1/N < MPCR < 1$ (see Focus 5.3 for details), it is individually rational to put the whole endowment in the private investment, although everyone in the group would be better off by favouring the public investment. The driving force of this result is free-riding: rational individuals only take into account the private return of their investment when considering the public investment. No matter what others do, it is payoff-improving to benefit from others' investment (if any) and invest everything in the private good and enjoy the private return from the private investment.

Isaac et al. (1984) is among the earliest attempts to experimentally investigate the empirical relevance of such behaviour (see, e.g. Ledyard, 1995, for a review of the literature). They consider repeated VCM games in which the same four players interact 10 times together. Two versions of the game are considered: one with low MPCR (equal to 0.3), another with a higher one, equal to 0.75. According to theory, the closer the MPCR

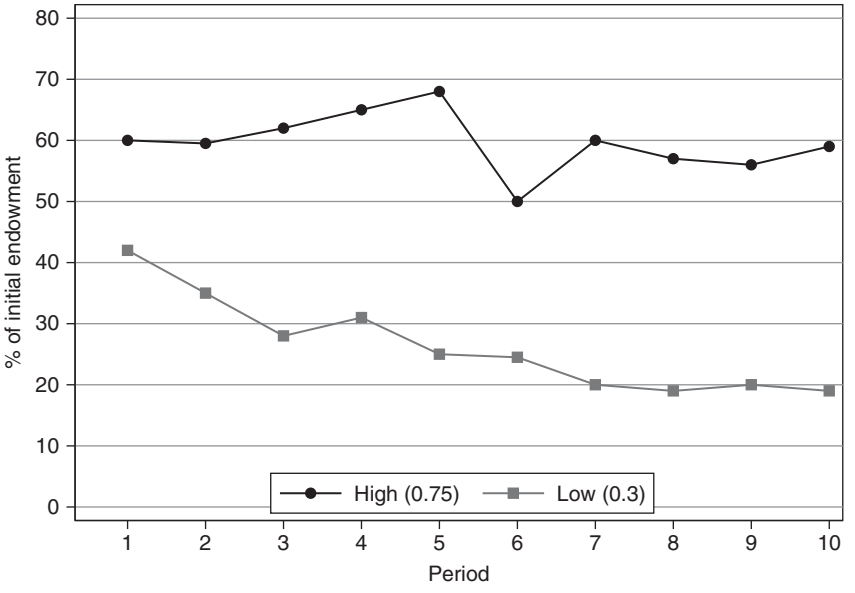


Figure 5.1 Empirical free riding in VCM games

Note. For each period in abscissa, the figure shows the average group contributions as a percentage of the optimal one (investing the whole endowment). Each line refers to a different treatment, with varying levels of the MPCR.

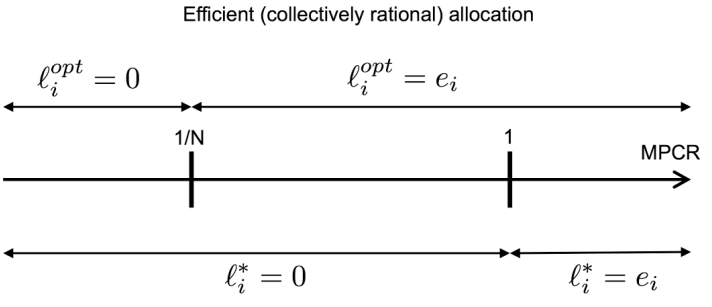
Source: Isaac et al. (1984, p. 135, Figure 4).

is to 0.25 ($= 1/N$ in the experiment), the stronger are the incentives to free-ride – conversely, cooperation becomes more and more likely as the MPCR becomes closer to 1, where individually rational behaviour spontaneously switches to the public investment. Several interesting lessons arise from the results, presented in Figure 5.1.

First focusing on the low-MPCR treatment, empirical behaviour clearly contrasts with the theoretical prediction: the contribution rates are strictly positive and amount to 40% to 20% of the initial endowment. It is worth noting that, although far from the Nash equilibrium, this behaviour is just as far from the fully cooperative outcome one would obtain if people cared about others just as much as they care about themselves. The pattern over time is also worth noting: the cooperation rate is decreasing over time, reaching its lowest level at the final stage of the experiment. This is a typical outcome in this kind of experiment, called an *end-game effect*. Overall, these results show that the free-riding issue in public-good-provision problems might well be weaker than expected. It does not rule out any explanatory power of theory, though. Turning to a comparison between the two treatments, it clearly appears also that insights about how behaviour changes according to the value of the MPCR are accurate. Contributions are much higher when the MPCR in higher (equivalently, contributions are much lower when incentives to free-ride are higher), and the decrease over time is also more attenuated. The general lesson from this seminal work is twofold: theory accurately describes how behaviour is adjusted to the monetary incentives at stake, but definitely misses something in driving forces of behaviour itself.

Focus 5.3
Equilibrium analysis of the VCM game

The theoretical analysis of the VCM aims to answer two different questions: what can small economy of N people best achieve given the available endowment? And what they will actually do if choices are not constrained – i.e. what allocation will result on a competitive market? Answering the first question amounts to comparing what can be collectively achieved according to the whole set of possible investments. Since the return rates all are linear, the answer is quite simple. One dollar from the endowment results in a wealth equal to q if invested in the private good, and equal to Q if invested in the public good. These are the public returns of the investment possibility, as they measure the overall change in wealth in the whole community associated with each possibility. The resources are thus best used by investing anything in the public good when $Q > q \Leftrightarrow Q/(qN) > 1/N$, and by opting for the private investment otherwise. This is the efficient-allocation rule (which corresponds to what is known as the Bowen–Lindahl–Samuelson condition in public economics). Let's now investigate how people will be individually willing to behave in this environment. Again, the linearity of the returns makes the problem straightforward. While deciding on the investment of each dollar from one's own endowment, each member of the group compares the private returns from the investment – by how much one's own wealth increases. As stated in the text, this return is still q for the private good, but is equal to Q/N for the public investment. The individually rational strategy is thus driven by the rate of substitution between the public good and the private good, usually called the *marginal per capita return* in the literature: $MPCR = Q/(qN)$. The private investment dominates the public one at the individual level if the MPCR is lower than 1 ($Q/N < q$). Each individual will then decide to opt for the private investment, resulting in equilibrium contributions $\ell_i^* = 0$, $\forall i$ (and $\ell_i^* = e_i$, $\forall i$ as soon as $MPCR > 1$). The discussion is summarised in the figure below.



Not all ranges of the parameters give rise to a market failure. If $MPCR < 1/N$ or $MPCR > 1$, individually rational decisions coincide with the efficient allocation and there is nothing to worry about. But if $1/N < MPCR < 1$, a social dilemma arises: individual decisions no longer match the efficient allocation, because individuals fail to take into account the consequences of their investment for the rest of the community. Each individual is better off opting out of the provision of the public good.

In terms of internal validity, the two kinds of result make use of different identifying assumptions. Comparisons between treatments only require that no confounding effect is correlated with the treatment – if the room temperature influences behaviour in a particular way, and both treatments have been implemented in the same experimental lab, then it is neutral on inferences based on behavioural variations. Inference about free-riding behaviour per se, by contrast, is a measurement problem. It requires that observed choices are induced by the chosen environment. If people rather react to features beyond the experimenter's control, then observed choices do not inform about target behaviour.

Andreoni (1995) designed an experiment aimed at addressing this second issue. The main research question is whether observed contributions from participants to VCM experiments are actually motivated by non-purely selfish preferences, like kindness or altruism. If not, in Andreoni's (1995 p. 893) own words, 'a second hypothesis is that experimenters have somehow failed to convey the incentives adequately to the subjects … subjects have somehow not grasped the true monetary incentives'. Andreoni labels this failure of internal validity 'confusion'. To that end, the experiment considers three treatments. The standard public-good game (treatment REGULAR) is similar to the ones considered above – groups are made of five subjects playing 10 VCM games with an MPCR of 0.5. The main difference is that subjects play with different others in each game. The main treatment of interest, labelled RANK, aims to eliminate any other-regarding motive while maintaining the same incentive structure of the game. In this treatment, subject's final earnings do not depend on their absolute earnings from the game, but rather on how their earnings compare to other subjects in their group. A list of fixed prizes is announced before the game takes place. In each period of play, the subject ranked first gets the highest pay-off, the subject ranked second gets the second-highest payoff and so forth. The incentive structure of the game remains the same: the dominant strategy, just like in the REGULAR treatment, is to contribute nothing to the public good. But the incentive to contribute due to other-regarding motives is now arguably eliminated from the game: since the public investment earns just as much for anybody in the group, contributing does not benefit others and just harms the investor's ranking. Positive contributions in this treatment thus cannot be interpreted as a departure from free riding due to kindness towards others.

Thanks to this feature of the RANK treatment, the comparison between these two treatments aims to identify the extent of contribution that is actually due to kindness – and what share of the usually observed level of cooperation can actually be attributed to confusion or error. There is, however, a potential confounder in this comparison. There are two actual changes in the decision environment between the two treatments. One is the change in the compensation scheme, which is implemented on purpose. But this requires a second change: the ranking information becomes available to subjects. If such information has an influence on behaviour (e.g. by fostering relative comparisons), there is no way to disentangle the two effects based on a simple comparison between RANK and REGULAR. A third treatment is designed to address this issue: the REGRANK treatment implements the same compensation rule as the REGULAR treatment, but provides subjects with the ranking information at the end of each round of play.

Table 5.1 Voluntary contributions without altruism

	Percentage of endowment contributed to the public good, by round										
Condition	1	2	3	4	5	6	7	8	9	10	All
REGULAR	56.0	59.8	55.2	49.6	48.1	41.0	36.0	35.1	33.4	26.5	44.07
REGRANK	45.8	45.4	32.6	25.0	23.1	17.8	11.3	9.5	8.3	9.0	22.79
RANK	32.7	20.3	17.7	9.9	9.2	6.9	8.1	8.3	7.1	5.4	12.55
$\neq$	13.2	25.1	15.0	15.1	13.9	11.0	3.2	1.3	1.2	3.6	10.24
% Regular	23.5	42.0	27.1	30.4	28.9	26.7	8.9	3.6	3.6	13.5	20.82
	Percentage of subjects contributing zero to the public good										
Condition	1	2	3	4	5	6	7	8	9	10	All
REGULAR	20	12.5	17.5	25	25	30	30	37.5	35	45	27.75
REGRANK	10	22.5	27.5	40	35	45	50	67.5	70	65	43.25
RANK	35	52.5	65	72.5	80	85	85	85	92.5	92.5	74.50

Note. For each round of play (in column) and in each treatment (in row) the upper part provides the average observed contribution as a percentage of total endowment. The lower part displays the share of subjects who behave as perfect free-riders – i.e. contribute exactly 0.
Source: Andreoni (1995, p. 896, Tables 1 and 2).

The results are shown in Table 5.1, providing both the average level of contribution in each round, and the share of subjects who behave as perfect free-riders. According to both outcomes, the three treatments are perfectly ordered: contributions are always higher in REGULAR as compared to REGRANK, and higher in REGRANK than in RANK. The change in behaviour from REGULAR to REGRANK confirms that informing subjects about their relative performance changes contributions – but it cannot be attributed with certainty to either kindness or confusion. By contrast, only confusion can explain positive contributions in the RANK condition. The average behaviour in this treatment, in terms of both contribution level and perfect free-rider distribution, amounts to half the one observed in REGRANK. This implies that half the contributions observed in this treatment (and almost one-third of those observed in REGULAR) have nothing to do with subjects' willingness to improve the group's outcome. This is only the empty part of the glass, of course; the full part is that half of observed contributions can be taken as accurate measures of people's tendency to spontaneously overcome the free-rider problem. This is large enough to deserve attention – and did elicit a large body of literature in the last three decades.

5.2 The Incentive Structure of Experiments

Smith's view of an experiment as a microeconomic system identifies the main building blocks of empirical identification based on experiments. Inference is based on outcome behaviour, supposedly generated by the environment and institutions implemented in the laboratory. The link between the two, which makes the whole work as a system, is individual behaviour: outcome behaviour results from decisions by people, in response

Table 5.2 Smith (1982) precepts: three incentive-compatibility criteria

Criterion	Description
Non-satiation	More is always better than less
Saliency	Payoff differences are such that choices are worth it
Dominance	The whole experiment is attractive enough to compensate for the opportunity cost of participation

to the experiment rules. There is one driving force in this system that remains beyond control: individual preferences over outcomes. Such a system is closed only if preferences driving decisions are well defined over outcomes. This is the main rationale for the use of monetary incentives in experimental economics – how much people lose or win according to what happens in the course of an experiment. These principles are described in the next subsection. As all methodological rules described in this chapter, this one has pros and cons and experiences famous exceptions. The most noticeable is even the field as a whole, as the use of incentives is to a large extent specific to economics among all experimental social sciences. Beyond this variety across fields, the choice of incentive structure raises practical issues within economics, which we describe in the last two sections.

5.2.1 The Logic of Incentives

The choice of the incentive structure of an experiment has been introduced as the core internal-validity issue in Smith's (1982) seminal article.[1] This choice is very much like a mechanism design problem: the incentive structure is what makes individual decisions driven by the environment and the institutions. Smith characterises the properties of the incentive structure according to three precepts – i.e. criteria, rather than rules – to assess the accuracy of the incentive structure. They are summarised in Table 5.2.

The first criterion is '*non-satiation*', which prescribes that more must always be better for everyone, and at any point, in the experiment. In case of a costless choice between two possible options, where the second is offering a higher reward, non-satiation thus requires this second alternative to be strictly preferred to the first one. The risk otherwise is to see people in the experiment not caring about the consequences of their own choices – possibly without any possibility to identify them. The application of this precept might seem very intuitive and almost overly obvious. One common limitation to non-satiation comes from threshold effects. For example, when exam grades are used as a rewarding currency, the reward increases to the extent the grade does. However, above a certain threshold, students might no longer care about additional gains. Giving a bonus

[1] In the words of Smith (1982, p. 935), non-satiation and saliency precepts are 'sufficient conditions for the existence of an experimental micro-economy, that is, motivated individuals acting within the framework of an institution'. With the addition of dominance (and privacy, to be discussed in Section 5.2.4), the experiment is a 'controlled microeconomic experiment [overcoming the possibility that] individuals may experience important subjective costs or values in transacting, and may bring invidious . . . taste to the laboratory from everyday social life'. In Smith's paper, a fifth precept is introduced, 'parallelism', which refers to external validity and will hence be discussed in Chapter 8 – Section 8.1 in particular.

of 20 points to all subjects gives rise to satiation for a student who already got 90/100, resulting in a flat compensation system once 10 additional units have been accumulated in the course of the experiment.

A second criterion is '*saliency*', meaning that the decisions in the experiment must be unambiguously linked with rewards. This implies that the differences in payoffs, or the marginal utility of the compensation scheme, must noticeably vary according to choices. For example, if one decision implies earning $2, and the next-best decision implies earning $2.25, it is not clear that for all agents this will be a sensible increase, as not even a cup of coffee can be bought with a quarter. Although the principle is that difference in level should be convincing, and make a difference for the decision-maker, there is no clearly and universally defined criterion to assess saliency. Illustration 5.3 provides an example of how sensitive outcome behaviour can be to the saliency of the experimental stakes.

A last criterion in the design of incentives is '*dominance*', which implies that the reward structure dominates any cost associated with participation in the experiment, both inside and outside the experiment. This criterion is akin to a participation constraint in microeconomic theory. Dominance implies a set of conditions on the compensation scheme used in the experiment. First, the compensation scheme must compensate the agent for the cognitive effort underlying decision-making in the experiment. Second, the compensation scheme must compensate the opportunity cost of participating in the experiment. The risk incurred in case of failure of the dominance principle is essentially a matter of selection bias and heterogeneity, as only those people for whom it is worth it will actually care about their decisions, or even come to the laboratory. As for all the precepts described here, there is no clear-cut reference that can be used to establish how much people should be compensated. It depends in particular on the market wage, the value of time and individual characteristics (such as human capital) of the subject pool, as they might all potentially affect the opportunity cost of individuals.

The three precepts together provide guidelines over the choice of the incentive structure of the experiment: how to design, in level and variation, the relationships between the states of the system and individual payoffs. In the terminology of the mechanism-design literature, compliance with the precepts makes the experiment incentive-compatible. Thanks to this property, the experiment offers control of preferences through the control of incentives.

5.2.2 Why Incentives after All?

Based on the above logic, the use of monetary incentives is almost systematic in experimental economics (at least to the extent that not using incentives must be strongly justified). Ortmann (2010), for instance, reports that all experimental studies published in the *American Economic Review* between 1970 and 2008 paid subjects according to their performance. It is, however, a matter of intense debate, both within economics and between economics and other experimental social sciences – psychology in particular. Within economics, one of the most convincing advocates against the use of monetary incentives is Ariel Rubinstein (2013, p. 541), who for instance notes,

Illustration 5.3

Saliency and coordination: experimental evidence based on the stag hunt game

The stag hunt game is a famous coordination game drawing back to the French philosopher Jean-Jacques Rousseau, who introduced it as a metaphor of collective action and social cooperation. Two players have the choice between hunting a stag or a hare. Hunting a stag (action *A*) is hard and requires the joint effort of both players, but then the reward is relatively high – large meal. Hunting a hare (action *B*) is easy and each player can succeed on their own, but then the reward is relatively low – small meal. The actions in this game strongly depend on beliefs about what the other player will do. The game has two symmetric equilibria: the outcome maximises payoff if both players hunt a stag (the outcome is said payoff-dominant), but if there is any doubt about what others will do, then hunting a hare is a riskless action (this outcome is hence the risk-dominant equilibrium of the game). Which of the two will be selected is an empirical question. Battalio et al. (2001) consider three variations of this game, presented below.

Game 2R

	A	*B*
A	45, 45	0, 35
B	35, 0	40, 40

Game R

	A	*B*
A	45, 45	0, 40
B	40, 0	20, 20

Game 0.6R

	A	*B*
A	45, 45	0, 42
B	42, 0	12, 12

While the strategic structures of all three instances are the same, the size of the incentives strongly differs. The *optimisation premium* – i.e. the difference between the payoff of the best response to an opponent's strategy and the inferior response – is twice as large in game 2R as it is in game R, and six-tenths as large in game 0.6R as it is in game R. The experimental implementation of these games aims to assess to what extent a change in incentives induces a change in coordination, based on three main theoretical hypotheses: (i) the larger the optimisation premium, the more responsive the subjects' behaviour will be to beliefs; (ii) the larger the optimisation premium (i.e. higher sensitivity to the history of one's opponent's play), the faster behaviour converges on an equilibrium; (iii) the smaller the optimisation premium, the more likely the behaviour converges on the payoff-dominant equilibrium (A, A). Eight cohorts of eight subjects are randomly paired and play one of the three games seventy-five times. Observed behaviour is summarised in the table below (from Battalio et al. 2001, p. 754, Tables 1 and 2).

	Period 1		Period 75	
	A	*B*	*A*	*B*
0.6R	41 (0.64)	23 (0.36)	28 (0.44)	36 (0.56)
R	45 (0.70)	19 (0.30)	16 (0.25)	48 (0.75)
2R	34 (0.53)	30 (0.47)	3 (0.05)	61 (0.95)
Total	120 (0.63)	72 (0.73)	47 (0.24)	145 (0.76)

Two important results emerge: while initial behaviour is the same across treatments, it converges on very different outcomes after some repetitions. After a while, participants play the

payoff-dominant action more often the larger the optimisation premium is. The study of the dynamics of behaviour between these two time periods provides support to the three above hypotheses. The general lesson from this experimental evidence is that the size of the stakes strongly influences strategic behaviour in the laboratory. The open question is then which of the observed behaviours is more informative about game-theoretical predictions.

I have never understood how the myth arose that paying a few dollars (with some probability) will more successfully induce real life behavior in a subject. I would say that the opposite is the case. Human beings generally have an excellent imagination and starting a question with 'Imagine that . . . ' achieves a degree of focus at least equal to that created by a small monetary incentive.

As a matter of fact, there are noticeable exceptions to this rule, which even elicited a large literature in behavioural economics. Illustration 5.4 provides an example of such a very influential laboratory experiment which does not make use of economic incentives – and only relies on fictitious scenarios.

This kind of counterargument echoes the view of incentives that dominate the experimental literature in psychology. This has been popularised, for instance, by Tversky and Kahneman (1986, p. 274), who conclude a survey on the topic by noting, 'Experimental findings provide little support to [the] view . . . that the observed failures of rational models are attributable to the cost of thinking and will thus be eliminated by proper incentives'. This methodological debate is still ongoing, and is likely to remain open for a long time. It amounts to an empirical question: is behaviour less or more conclusive as an empirical outcome of the experiment system when performance is incentivised? Several attempts to answer the question have been made in recent years, of which Illustration 5.5 provides an example. Two different cases have to be distinguished in order to make a choice in this regard. The first question is whether or not the use of incentives harms inference based on experimental behaviour. Although a few examples go in this direction, there is very little evidence supporting this (other) view (as well). On the other extreme, it is rather clear that not incentivising performance leads to more noise in the data (Smith and Walker, 1993). Since decisions of the agents no longer have monetary consequences for them, the motivations behind individual behaviour are more likely to be idiosyncratic and diverse. In the extreme case of surveys based on purely declarative answers, the resulting measures will likely not be very informative about the underlying true attitude. If one asks subjects whether they feel happy in the experiment, the observed level of happiness will be regarded as a poor measure of the true mental state of the subjects by most empirical economists. Any systematic change in behaviour, however, like the difference in self-reported happiness between two treatments, is hardly induced by such noise and provides convincing cues about the actual happiness effect of the treatments. This again illustrates that the choice of incentivising behaviour has to be mainly driven by the research question. More generally, it also points to the answer, giving rise to the current consensus in the experimental economics community. If (when) incentives do not harm, they at worst are innocuous on behaviour and at best enhance

Illustration 5.4

Evidence from non-incentivised behaviour: the status quo effect

The *status quo bias* is another striking behavioural consequence of prospect theory presented in Focus 5.2. One of the earliest empirical studies documenting this phenomenon is due to Samuelson and Zeckhauser (1988), based on two simple surveys. The first survey tells the respondent: *You are a serious reader of the financial pages but until recently have had few funds to invest. That is when you inherited a large sum of money from your great-uncle. You are considering different portfolios. Your choices are*:

___ a Invest in moderate-risk Co. A over a year's time; the stock has 0.5 chance of increasing 30% in value, a 0.5 chance of being unchanged, and a 0.3 chance of declining 20% in value.

___ b Invest in high-risk Co. B over a year's time; the stock has a 0.4 chance of doubling in value, a 0.3 chance of being unchanged, and a 0.3 chance of declining 40% in value.

___ c Invest in treasury bills. Over a year's time, these will yield a nearly certain return of 9%.

___ d Invest in municipal bonds. Over a year's time, they will yield a tax-free return of 6%.

In the second questionnaire, the text includes a slight modification with respect to the initial conditions: '*You are a serious reader of the financial pages but until recently have had few funds to invest. That is when you inherited a portfolio of cash and securities from your great-uncle. A significant portion of this portfolio is invested in a moderate-risk Company A (option (a)). You are deliberating whether to leave the portfolio intact or to change it by investing in other securities (the tax and broker commission consequences of any change are insignificant).* The proposed choices are identical to the one shown along with the first questionnaire. In one case (the second survey) the money has already been invested, whereas in the other case (the first survey), the money has not been invested yet. Obviously, this leaves unchanged the comparison between all options and the instructions aim to make clear that changes to the portfolio induce no monetary cost. While economic theory imposes no condition about the choice between the four options (this entirely depends on the shape of individual preferences), it should, however, be the case that the arbitrage decisions of the subjects in the two situations are the same. Observed behaviour strongly contradicts this prediction, with an average 20% more choices of the default option in the second survey. This is consistent with loss aversion deduced from prospect theory – the default investment working as a reference point from which departures are evaluated in the loss domain. Alternative explanations have been raised in the literature, such as costs of thinking, small transaction costs or psychological commitments to prior choices.

the quality of inferences based on behaviour. So just use it, for the same outcome if not for a better one.

5.2.3 Implementation Issues: Multiple Play Incentives

In many experiments, the decision task is repeated in an effort to allow for some learning and to avoid focusing only on initial responses (see Section 5.5 for a discussion about

Illustration 5.5
The effect of incentives on experimental outcomes

Camerer and Hogarth (1999) review 74 studies published in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy* and the *Quarterly Journal of Economics* between 1990 and 1998. Studies were included if they compared behaviour of subjects according to their performance with different levels of incentives ranging from no incentives to high monetary incentives. Camerer and Hogarth (1999) classify the studies in three broad classes depending on the effect of incentives. In the first class of studies, incentives help improve performance in experimental tasks. This appears to be the case for judgement and decision tasks: incentives promote effort in memory and recall tasks of past events, as well as increasing attention. Incentives are important to increase effort in mundane clerical tasks (coding words or numbers, building things). In this class, the main effect on effort is obtained by raising incentives from hypothetical choice to incentivised choice. Increasing the level of incentives appears to have limited effect. In the second class of studies, incentives do not appear to matter much because the marginal return on effort is low. According to Camerer and Hogarth (1999), this is the most common result. This class regroups studies in experimental games, auctions and preference elicitation. The marginal return on effort is low when it is hard to improve performance (computing all the equilibrium strategies of a game) or when performance is easy to attain (when the strategy is obvious to most participants). If incentive does not improve or hurt average performance they decrease the variation in performance, a point raised early by Fiorina and Plott (1978). In the third class, incentives do hurt performance. This class is the smallest one and regroups mostly judgement and decision tasks. Incentives hurt here because of a number of reasons: they push subjects to stick to a given heuristic, they make subjects overreact to feedback, and they make participants self-conscious about tasks which should be automatic. The general lesson thus is that incentives generally seem a good idea to get data of better quality, but different research questions and researcher belief about the true (behavioural) data-generating process might well lead to different choices.

the implementation of repetition itself). The same players are then involved in the same task (either decisional or strategic) several times. The general principles behind the use of incentives applies to each one of these multiple decisions. In this case, subjects would earn some money based on their performance at each stage of the repetition, and their overall compensation for participating in the experiment would be computed as the sum of their earnings over all instances of the decision task. Such a compensation scheme, however, raises important internal-validity issues.

When subjects face real play in several successive tasks, the outcome obtained in the earlier tasks can contaminate behaviour in the subsequent tasks and lead to biased measurement. Several well-documented phenomena can give rise to such carry-over or contamination effects: wealth effects, house-money effects and portfolio effects. *Wealth effects* are the most obvious consequence of paying subjects for the sum of their earnings in all decision tasks. As the experiment evolves towards subsequent decision stages, the

level of wealth of the subjects increases thanks to accumulated earnings (either known or expected) at each stage. If wealth has an effect on decision-making (income effects in the utility function are an obvious reason for this to occur), then decisions at later stages are not similar to decisions made earlier in the experiment: there is serial correlation in decisions due to the design of performance-based incentives. Such a wealth effect is also likely to occur in decision experiments involving uncertainty as soon as risk aversion changes with wealth. When subjects have decreasing risk aversion, experiencing prior gains increases wealth and potentially leads to higher risk taking in subsequent tasks. On the other hand, it could also be the case that subjects show a propensity to break even after a prior loss and take more risks as the experiment progresses. Thaler and Johnson (1990) show evidence of such an effect of prior gains on risk behaviour, and label *house-money effects* biases arising from the fact that subjects consider prior outcomes as windfall money and take more risk with it.[2] *Portfolio effects* come from the fact that changing behaviour provides a natural hedge in experiments where uncertainty plays a role. For example, in an experiment involving an unknown urn filled with two balls of different colours, taking two complementary positions on and against a colour in two choices provides a hedge against uncertainty.

Because of these likely failures of the internal validity of multiple decisions, the implementation of incentives is often adapted accordingly. One of the most widely used compensation schemes to circumvent these caveats is the so-called *random incentive system* (RIS), which amounts to paying for real only one of these tasks, chosen at random, at the end of the experiment. The main advantage of such a system is to isolate, through randomisation, one choice from another. The intuition is rather straightforward: if the experiment involves two decision tasks, each one compensated with a one-half probability, then an expected utility maximiser will put exactly the same weight on the two outcomes. Focus 5.4 summarises the main methodological drawbacks of this procedure.

Another drawback of random incentives lies in their saliency. Randomisation decreases the expected value of the incentives, as each task is paid with a probability lower than 1. The size of the stakes thus has to be adjusted accordingly. Conversely, for this same reason, RIS also allows us to study decision tasks with big monetary consequences at a reasonable cost. This need to adjust compensation for decreased saliency is shared by all random incentive compensation schemes. It is even more stringent for between-subjects RIS, in which one decision for only one player out of all participants is compensated. In an experiment involving J decisions and N participants, pure between-subject randomisation induces a $1/(NJ)$ probability that each decision actually counts in terms of payoff. On top of this saliency issue, a violation of dominance can also potentially arise when using this mechanism if subjects who happen not to be paid *ex post*

[2] A possibility to mitigate this income effect is to postpone the disclosure of the draws in the chosen lotteries to the end of the experiment. Another possibility, introduced by Holt and Laury (2002), is to require the subjects to give up previous gains in order to answer to subsequent tasks. However, in practice, any perceived change in the expected value of the experiment can influence risk attitudes, as stressed by e.g. Grether and Plott (1979).

Focus 5.4
Incentive-compatible compensation of repeated choices: the random incentive system

Baltussen et al. (2012) list at least five different names for the random incentive system described in the text, among them 'random lottery incentive system', 'random lottery selection method', 'random problem selection procedure' and 'random round payoff mechanism'. This incentive scheme blocks the changes in wealth over the course of the experiment. According to Holt (1986), however, subjects might consider the experiment a meta-lottery where each task can be selected with equal probability. More generally, subjects can consider the experiment a meta-lottery with any probability distribution over the different tasks depending, for example on the precise form of the random incentive system or on their beliefs. As a consequence, subjects might no longer perceive each task in isolation and integrate all the choices in the meta-lottery, leading to carry-over effects similar to the ones identified when all tasks are paid. Because each task corresponds to a given outcome in a (meta-)lottery, these complementarities exist when the independence axiom for choice under risk is violated. Several studies have investigated the internal validity of the random incentive system. Bardsley et al. (2010, p. 269), show that the preceding speculations are incorrect. In their words, 'It is easy to see, however, that the RLI [RIS] could be unbiased in the presence of any form of non-EU preferences given different assumptions about how agents mentally process tasks.' Starmer and Sugden (1991) show that these potential problems are of little concern and that isolation can be assumed. A large body of literature has confirmed this finding (Cubitt et al., 1998; Hey and Lee, 2005; Lee, 2008; Baltussen et al., 2012). The overall picture is that for simple binary choices under risk with a high number of repeated measurements, the random incentive system is compatible with the isolation hypothesis. In more complex tasks or dynamic tasks, or in case of between-subject randomisation, however, the existing evidence shows that isolation might not be as strong and carry-over effects can appear (see Beattie and Loomes, 1997; Baltussen et al., 2012, for further details). Cox et al. (2015) provide an empirical investigation of the incentive properties of a wide variety of compensation mechanisms.

see their participation as a pure waste of time. The anticipation of such a feeling in the course of the experiment might undermine the ability of incentives to compensate the opportunity cost of taking part in the experiment.

5.2.4 Other-Regarding Preferences and the Incentive Compatibility of Experiments

The logic behind the use of monetary incentives is to provide control of subjects' preferences over outcomes in such a way that the experimental situation implements an actual microeconomic system. The experiment then provides evidence on the outcomes raised by a given combination of the environment and the institutions. This aims to mimic the way theory works, in which preferences are given. Accumulated evidence over the last decades, however, tends to challenge the idea that individual monetary payoffs are enough to describe preferences (see e.g., the discussion associated with the prisoners' dilemma game in Section 1.3.1 or the VCM game in Section 5.1.4). Rather, in many circumstances, people seem to behave differently according to the consequences

of their decisions for others. Such motives are often labelled other-regarding preferences to point out the departure from the standard self-interested representation of individual preferences. Illustration 5.6 describes the ultimatum-bargaining game, which contributed to stimulating interest in this topic.

As an illustrative example, consider an outcome-based model of social preferences (Focus 5.5 describes an example from the alternative class of intention-based models). Individual utility, U_i, is defined over two attributes: one's own payoff, x_i, and other players' payoff, x_{-i}. Various representations of the dependency of agent i utility on x_{-i} exists in the literature. If the utility is independent of x_{-i}, the model coincides with the standard self-interest assumption. Alternatively, the utility can be defined over the total surplus ($\sum_j x_j$), so that social preferences mimic a utilitarian social planner. In Bolton and Ockenfels (2000), utility is a function of the share of agent i in the allocation ($x_i / \sum_j x_j$). In Charness and Rabin (2002) utility is a function of a disinterested social-welfare criterion, i.e. a weighted sum of the total surplus and of the payoff of the least well-off agent ($min\{x_i, x_{-i}\}$).

One of the most widely used outcome-based models is the aversion-to-inequality model introduced by Fehr and Schmidt (1999), in which utility is a function of payoff differences between the agents. In this model, the utility function is defined over the vector of individual monetary payoffs according to:

$$U_i(x_i, x_{-i}) = x_i - \underbrace{\theta_i^- \frac{1}{N-1} \sum_{j \neq i} \max\{x_j - x_i, 0\}}_{\text{loss from disadvantageous inequality}} - \underbrace{\theta_i^+ \frac{1}{N-1} \sum_{j \neq i} \max\{x_i - x_j, 0\}}_{\text{loss from advantageous inequality}}$$

Each agent thus has three sources of utility: one's own individual payoff and two functions of payoff differences leading to utility losses in case of both disadvantageous inequality ($x_j > x_i$) and advantageous inequality ($x_j < x_i$). The original model adds three assumptions to the parameters value: (i) agents suffer more from disadvantageous inequality than from advantageous inequality: $\theta_i^+ \leq \theta_i^-$, (ii) agents do not like advantageous inequality: $\theta_i^+ \geq 0$, and (iii) no agent is willing to burn money in order to reduce inequality: $\theta_i^+ < 1$.[3] Under this set of assumptions, agents endowed with such preferences exhibit aversion to inequality: there is a trade-off between one's own payoff and the fairness of the resulting allocation.

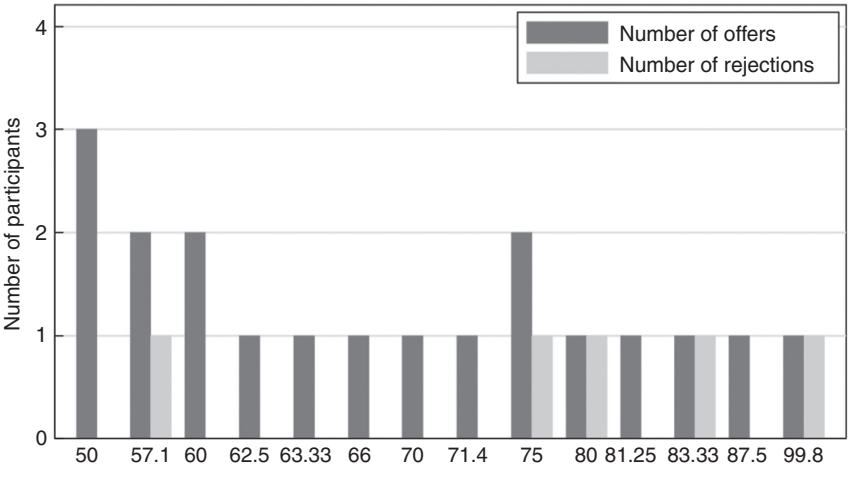
The main point as regards performance-based incentives in experiments is that if subjects exhibit this kind of preference, then part of the control implemented through the use of monetary incentives is lost. As an example, consider a prisoners' dilemma game. Table 5.3.a shows the typical payoff matrix of the game, in which each player can choose between two actions: cooperate (Coop) or defect (Def). As discussed in Section 1.3.1, empirical behaviour often does not coincide with the Nash equilibrium of this game: while defecting is the individually rational action, many people decide to cooperate. This is a departure from the Nash equilibrium when payoff accurately describes individual preferences. But for inequity-averse subjects, the game played is actually the one described in Table 5.3.b. For such a payoff structure, cooperating is a

[3] The model embeds other kinds of psychological motive with alternative parameterisation: individual preferences exhibit guilt if $\theta_i^+ < 0$ and is envious if $\theta_i^- < 0$.

Illustration 5.6

Social preferences and strategic uncertainty: the ultimatum-bargaining game

The ultimatum-bargaining game (UBG) introduced by Guth et al. (1982) is an early experiment that stimulated research into social preferences. It focuses on a simple two-player game. One player is the sender, the other is the receiver. The sender receives an initial endowment and is asked how much is sent to the receiver. The receiver then decides whether to accept or reject the offer of the sender. The offer is implemented if the receiver accepts it, and each player gets the corresponding payoff. But if the receiver rejects the offer, both players get 0. The sub-game perfect equilibrium is rather simple: the receiver should accept any positive offer, leading the sender to offer the smallest possible share of the endowment. The game thus replicates a situation in which the sender has full bargaining power. The figure below (from Guth et al., 1982, p. 375, Table 5) shows the number of experimental subjects who decide to offer or reject the share of the endowment shown in the *x*-axis.



First, it is common for receivers to reject offers below 20% of the endowment. Moreover, most offers from senders lie in the [40%, 50%] interval, with no offers above 50% and very few below 20%. There are thus two deviations from the behaviour predicted by sub-game perfectness. While pure selfishness can hardly explain receivers' behaviour, one can wonder whether senders' behaviour is induced by other-regarding concerns or a best reply to rejection behaviour. This early evidence has been replicated many times since the appearance of the paper, with always the same patterns. In particular, this result appears robust to the level of incentives: in an experiment that raises the stakes to three times the monthly expenditure of the average participant, Camerer and Hogarth (1999) replicate the evidence on proposal rates.

Focus 5.5
Intention-based social-preference models

The altruism-based model has been later generalised to highlight intention-based social preferences. In such a model, what matters to an agent is not only the payoffs but also the intention behind others' actions. In Rabin's (1993) model of intention-based reciprocity, it is for instance assumed that people want to be nice to those who are nice to them, and punish those who are mean to them. In order to account for such behavioural motives, the standard game-theoretic approach must be generalised. To that end, Geanakoplos et al. (1989) introduce psychological games, in which payoffs depend not only on actions, as in traditional game theory, but also on beliefs about actions. Rabin (1993) applies these ideas in the simple context of a two-player game (later on generalised as a sequential game by Dufwenberg and Kirchsteiger, 2004). Let a_i denote the strategy chosen by Player i in their action set, b_{ij} denote Player i's belief about the strategy Player j is choosing and c_{ij} denote Player i's belief about what Player j believes Player i's strategy is; i.e. a_i are standard strategies, b_{ij} are first-order beliefs and c_{ij} are second-order beliefs. The kindness of Player i towards Player j is the difference between the material payoff to Player j, x_j, minus the equitable, or fair, payoff to Player j. The fair payoff to j is the average payoff of the lowest and the highest payoff Player i could have secured to Player j. The fair payoff corresponds to an average point on the Pareto frontier for Player j's payoffs and serves as a reference point to measure how generous Player i is to Player j. Formally, the kindness of Player i towards Player j is a function of both Player i's strategy, a_i, and the belief b_{ij} about Player j's strategy:

$$k_{ij}(a_i, b_{ij}) = \underbrace{x_j(a_i, b_{ij})}_{\text{payoff to Player } j} - \underbrace{\frac{1}{2}[\max_{\mathfrak{a}_i \in A_i} x_j(\mathfrak{a}_i, b_{ij}) + \min_{\mathfrak{a}_i \in A_i} x_j(\mathfrak{a}_i, b_{ij})]}_{\text{fair payoff}}$$

This function is equal to 0 if Player i's strategy gives Player j their fair payoff, negative if Player i's strategy gives Player j less than their fair payoff, and positive if Player i's strategy gives Player j more than their fair payoff (if possible). Player i also has beliefs about how kind Player j is to him. This belief function depends on first-order beliefs b_{ij} and on second-order beliefs c_{ij}. Formally, Player i's belief about Player j's kindness towards him is

$$h_{ji}(b_{ij}, c_{ij}) = \underbrace{x_i(b_{ij}, c_{ij})}_{\text{payoff to Player } i} - \underbrace{\frac{1}{2}[\max_{\mathfrak{b}_j \in A_j} x_i(\mathfrak{b}_j, c_{ij}) + \min_{\mathfrak{b}_j \in A_j} x_i(\mathfrak{b}_j, c_{ij})]}_{\text{fair payoff}}$$

The sign of the belief function reflects Player i's opinion about Player j's behaviour. For example, it is negative if Player i believes Player j is treating him badly. Utility functions driving behaviour in this context are augmented with kindness functions:

$$U_i(a_i, b_{ij}, c_{ij}) = \underbrace{x_i(a_i, b_{ij})}_{\text{material well-being}} + \underbrace{\theta_i k_{ij}(a_i, b_{ij}) h_{ij}(a_i, b_{ij})}_{\text{fairness}}$$

where θ measures how sensitive agent i is to reciprocity towards agent j. For each player, the strategies depend not only on material payoffs but also on beliefs about the other subject's intentions. For instance, if Player i believes that Player j is treating him kindly, then $h_{ij}(a_i, b_{ij}) > 0$ so that Player i will choose a strategy a_i such that $k_{ij}(a_i, b_{ij}) > 0$. Conversely, if Player i believes that Player j is treating him badly, then $h_{ij}(a_i, b_{ij}) < 0$ and Player i will choose a strategy a_i such that $k_{ij}(a_i, b_{ij}) < 0$. Rabin (1993) shows that there exists a 'fairness equilibrium' such that each player maximises utility and intentions are self-fulfilling and compatible ($a_i = b_{ij} = c_{ij}$). It can also be shown that fairness equilibria in games include Nash equilibria as specific cases where players mutually maximise or minimise each other's material payoffs.

Table 5.3 Outcome-based social preferences in the prisoners' dilemma game

	(a) Standard game		(b) Game played by inequity-averse subjects	
	Coop	Def	Coop	Def
Coop	5,5	–10, 10	5, 5	$-10-20\theta_1^-, 10-20\theta_2^+$
Def	10, –10	–5, –5	$10-20\theta_1^+, -10-20\theta_2^-$	–5, –5

Nash equilibrium if $\theta_i^+ > 1/4$.[4] As a result, there is a discrepancy between the game subjects are actually playing and the one analysed by the theoretical model – as illustrated by Illustration 5.7, the same kind of issue arises if one assumes alternative sources of other-regarding concern, like altruism. This has important consequences for the conclusions that can be drawn from such data: one cannot test at the same time theoretical assumptions about preferences and theoretical predictions about strategic interaction. If preferences are the main focus of the experiment, their occurrence does not challenge internal validity, provided their source is accurately controlled. Illustration 5.8 provides an example of such an experiment, which tries to disentangle intention- and outcome-based social preferences. But non-monetary motives will be confounding if behaviour is to be interpreted as conditional on the control over preferences offered by monetary incentives.

Beyond the case of social preferences in games, this feature has many consequences in the way experiments are designed and implemented. The general aim is to design and implement the incentive scheme of the experiment in order to minimise the confounding effects of any uncontrolled non-monetary motive of behaviour. First, this is the main reason why compensation rules and payoffs are described aloud publicly in experiments. This makes the incentive scheme common knowledge among participants in the experiment and prevents subjects forming beliefs about how well or badly they are treated as compared to others. Second, it is also sometimes useful to rescale the payoffs associated with decisions in the experiment using an abstract currency. Subjects then play with tokens rather than actual money, with possibly absolute values that are much higher

[4] Similarly, in the ultimatum-bargaining game presented in Illustration 5.6, observed behaviour is compatible with preferences such that $\theta^+ < 0.5$ for the sender and $\theta^- > 1/3$ for the receiver. If $\theta^+ > 0.5$, the sender always offers 50%, which is always accepted.

Illustration 5.7
Altruism in the prisoners' dilemma game

Altruism-based approaches have been developed as an alternative to outcome-based models. They assume that people care about the well-being of others. In its simplest, two-agent, version, utility has the following additive representation: $U_i = u(x_i) + \theta v(x_j)$. Here the total utility of the agent involves two components: the selfish part of utility, $u(x_i)$, is derived from one's own payoff, but, due to altruism, agent i also cares about the payoff of agent j. This is reflected by the second term, $\theta v(x_j)$. For the sake of simplicity, assume that utility is linear for both the selfish term and the altruistic term. In the prisoners' dilemma game described in Table 5.3.a, the actual game played by altruistic agents when faced with the monetary rewards of their actions is therefore:

	Left	Right
Top	$5 + 5\theta_1, 5 + 5\theta_2$	$-10 + 10\theta_1, 10 - 10\theta_2$
Bottom	$10 - 10\theta_1, -10 + 10\theta_2$	$-5 - 5\theta_1, -5 - 5\theta_2$

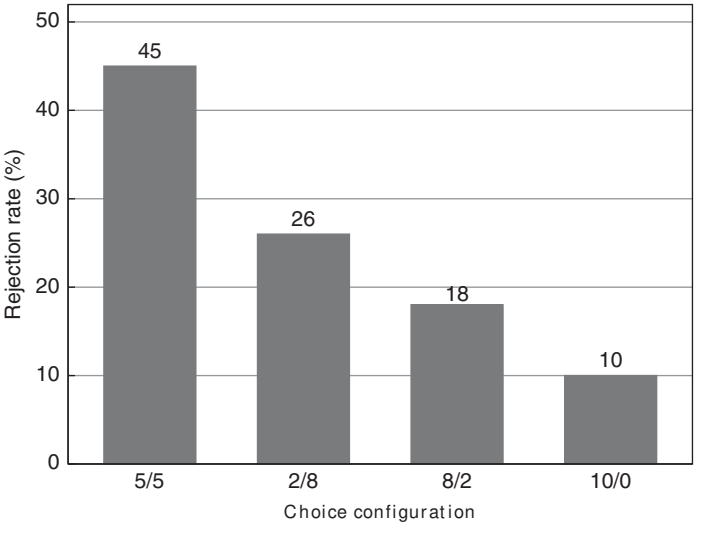
As for the inequity-averse example in Table 5.3.b, the incentive structure no longer provides perfect control of individual preferences – social preferences induce a discrepancy between the game actually played and the one described by monetary incentives. In this example, cooperation is an equilibrium if both players have a high enough level of altruism and discount the other rewards as at least one-third of their own rewards (*i.e*, if $\theta > 1/3$).

than their monetary equivalent. The monetary value of such '*experimental currency unit*' used to measure the payoffs in the experiment is generally announced before the start of the experiment, and gains are converted when subjects are informed about their overall performance. This is sometimes used to inflate the payoff, in order to improve saliency thanks to an illusion that amounts at stake are big even if they are associated with rather small monetary amounts. This is also useful if one needs to vary the monetary incentives across subjects or across implementations of the experiment (in the course of cross-cultural comparison, for instance; see Chapter 8, Section 8.3.3) while maintaining the currency used to label earnings in the experiment constant. Third, individual earnings from the experiment are confidential and private information of the owner. This avoids interpersonal comparisons, the anticipation of which might induce unwarranted competition between subjects during the experiment. Fourth, payment generally occurs in a separate room, where subjects enter individually to receive their payment. Fifth, this is also the reason why roles assigned to subjects are generally held constant when the experiment features roles that are not symmetric – i.e. a receiver and a sender. The main rationale is to avoid empathy, i.e. that a subject better cares about the partner's situation if playing in this position occurred earlier, or will occur later, in the course of the experiment. In a UBG game, for instance, it is likely people will behave differently depending whether they always decide as a sender, or are forced to put themselves in the shoes of the receiver by knowing they will be one at some point, or have been one before.

Illustration 5.8

Outcome versus intention: an experiment on the nature of social preferences

In order to disentangle outcome-and intention-based social preferences, Falk et al. (2003) design an experiment in which a given choice–outcome combination is associated with varying intentions. Subjects play a UBG with 10 tokens to split. Instead of implementing the usual choice of offers inside the range of all possible values, the experiment elicits a series of four binary choices. In all choices, one option is to divide the 10 tokens by keeping eight and leaving two to the receiver – allocation $(8, 2)$. Four possible alternatives are considered in turn: $(5, 5), (2, 8), (8, 2)$ or $(10, 0)$. Clearly, different behavioural models lead to different predictions of receivers' behaviour: a standard model involving self-interested agents predicts that the allocation $(8, 2)$ is never rejected by receivers; an outcome-based model (e.g. Fehr–Schmidt or Bolton–Ockenfels) predicts rejection of the unequal allocation $(8, 2)$ due to aversion to inequity whatever the alternative choice among the four is considered; last, an intention-based model predicts different rejection rates of the $(8, 2)$ allocation, depending on the alternative against which it has been chosen. The figure below (from Falk et al., 2003, p. 24, Figure 2) shows the observed rejection rates of the $(8, 2)$ offered in the four different choice configurations.



Overall, 55% of receivers reject the $(8, 2)$ allocation if the alternative share is $(5, 5)$, but only 10% do the same against a $(10, 0)$ alternative. Such a shift unambiguously supports the position that intentions matter in rejection behaviour. There is also evidence of pure aversion to inequitable shares, since 18% of the receivers reject the $(8, 2)$ allocation when $(8, 2)$ is the alternative – in this configuration the sender has no choice and cannot signal any intention.

These results provide evidence of outcome-based reasoning which contradicts a pure intention-based model. The general conclusion drawn by Falk et al. (2003) is that social preferences in this game are mainly driven by intentions, but influenced by outcomes. The

experiment also provides evidence that senders rationally respond to empirical rejection behaviour. As shown in the table below (from Falk et al., 2003, p. 24, Figure 1), the likelihood of the (8, 2) offer strictly increases in its expected return according to the empirical probability of acceptance.

Game	Expected payoff of the (8, 2) offer	Expected payoff of the alternative offer	Percentage of (8, 2) proposals
(5, 5) game	4.44	5.00	31
(2, 8) game	5.87	1.96	73
(10, 0) game	7.29	1.11	100

It is a well-document fact that roles interact with decision-making (see Illustration 5.9 for one of the most famous examples in social psychology). Finally, the nature of decision outcomes can be modified to control for unobserved preferences. For instance, Roth and Malouf (1979) use binary lotteries in bargaining games to control for risk aversion. With binary lotteries, the outcomes associated with decisions are lottery tickets, and payoffs are chances to obtain a fixed prize. If subjects are expected-utility maximisers and therefore reduce compound lotteries, the utility associated with the fixed prize is equal to the probability of winning that prize. For example, in an experimental bargaining game, the utility associated with an agreement would be equal to the percentage of lottery tickets received from this agreement (Murnighan et al., 1988). Because elicited probabilities capture von Neuman and Morgenstern utilities, binary lotteries are often used as devices to induce risk neutrality. Krawczyk and Le Lec (2015) show this design can also be used to induce more selfish behaviour in experimental games.

5.3 Parameters and Experimental Treatments

The use of performance-based compensation is a way to ensure the incentive compatibility of the experiment, i.e. that the chosen combination of the environment and institutions does induce individual behaviour. We now turn to the specification of these environment and institutions, and how it relates to internal validity. An important distinction needs to be made among the features of the system that are directly controlled by the experimenter. Some of them, called '(control) parameters', are set at the same value in all instances of the experiment. Others, called 'treatment variables', are purposefully varied across several instances of the experiment. The aim of this variation is to measure its effect on outcome behaviour. In order to provide proper identification of this effect, implementation of the treatment relies on the identification strategies described in Section 5.1.3.

Illustration 5.9

The effect of roles on behaviour: the Stanford prison (aborted) experiment

In a famous experiment, Stanford University psychologist Philip Zimbardo showed how far the disconnection of social and moral values can go when people are involved in alienating contexts – social positions driving decisions more strongly than individual preferences. The experiment took place in 1971 in the basement of Stanford University, which was equipped as a prison facility (the full story of the experiment is reported in Zimbardo, 2007). The participants were 24 students who responded to advertisements in local newspapers. The selected candidates were all male, mostly-middle class white, and were chosen conditional on a strong mental and emotional stability. The volunteers, who were all paid $15 per day for the experiment (which was supposed to last two weeks), were split into two groups: 12 were asked to play the role of 'guards', the other half were 'prisoners'. The guards were dressed in military uniforms, wore mirrored sunglasses to hide their eyes and carried wooden bats to intimidate the prisoners. No physical punishment was allowed but this was basically the only rule: for the rest, the guards were free to rule the 'prison' according to their own judgement. Conversely, the prisoners lived in tough conditions. Only cheap coveralls, very basic facilities and plain food were provided. A chain around their ankle was there to underline their status. Only identity numbers were used, instead of names. Upon students' agreement and thanks to the support of the local police, prisoners started the experiment without any warning, with a simulated raid of the policemen in their homes and a routine for real suspects, including fingerprinting. Zimbardo himself acted as a prison warden, so that he was able to directly observe the course of the study. The experiment degenerated very quickly and was suspended by Zimbardo after only six days. On one hand, guards quickly got carried away with a cruel authority, using punishments (mattresses were confiscated, access to the toilet was arbitrarily denied, etc.) and different kinds of humiliation. Overall, one-third of the guards began to show clear signs of sadism. On the other hand, many of the prisoners started to show symptoms of emotional and mental distress. Two of them were removed before the end of the experiment. Although a riot arose on the second day, none of the prisoners decided to leave the experiment before its end. Guards developed antisocial attitudes and showed no sympathy for individual protests. The main conclusion drawn from Zimbardo is that people who are drawn into some particular situations tend to adapt their behaviour to a role, instead of using their own judgement and moral values. Social and ideological factors shape this role, and consequently submerge individual personalities.

5.3.1 Direct Controls

All dimensions of the experiment that can be chosen by design are parameters of the experiment (those are the 'controlled inputs' of the framework developed in Chapter 4). Examples include, among many (many) others, the nature of the subject pool who will be invited to come to the laboratory, the number of periods of play, the exchange rate of the experimental currency, the number of subjects sitting in each section. Choosing the value taken by each such parameter in the experiment amounts to choosing the experimental DGP, hence providing direct control over the environment. Any different

choice in the value of one parameter is likely to change the outcome behaviour. For instance, designing an experiment implies making a choice as regards the size of incentives. Too low incentives may fail to implement salient enough decisions, while too high ones may mistakenly stimulate pay-off maximising behaviour (see Illustration 5.2 for an example). Whether the appropriate choice is closer to one extreme or the other has to be decided by sound judgement – according to a trade-off as regards internal validity which, in this case, also has to be weighted with the consequences in terms of the cost of the experiment.

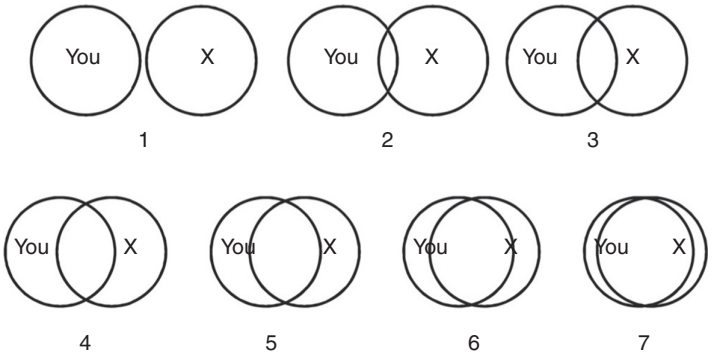
The experimental setting widens the set of dimensions that can actually be chosen. Induced-value designs (Smith, 1976), in particular, allow preference parameters to be chosen thanks to the incentive scheme. As shown in the auction example of Chapter 2, for instance, the marginal utility of subjects for the experimental good can be 'induced' by setting the price at which a unit of the good can be sold to the experimenter. Similarly, experiments focusing on effort at work often induce the marginal cost of effort (see Illustration 4.4 for an example): in this context, the effort at work is implemented as a simple number with positive monetary consequences for the employer and negative ones for the employee. In a non-laboratory context, these preference parameters are not only private and unobservable information, but also heterogeneous between subjects. Laboratory experiments, by contrast, allow one to include these dimensions in the set of chosen parameters.

Beyond the direct choice of the components of the decision-making environment, the design of the experiment also involves the set of measurement tools used to measure the existing heterogeneity leading to the outcome variable. This includes simple questionnaires collecting data on subject's socio-demographic characteristics like age, gender, field of study, etc. To this basic information, more specialised questionnaires can be added to measure specific dimensions of subjects' personality traits, morality or values (typically, based on questionnaires developed in psychology, see e.g. Borghans et al., 2008) or their cognitive skills – using, e.g., the cognitive reflection test (Frederick, 2005) or Raven's progressive matrices test (Raven, 2008). Illustration 5.10 provides an example of a measure of intensity of social relationships taken from social psychology. A recent trend in experimental economics also includes measurements of the physiological process and consequences of decision-making. For example, skin-conductance responses allow us to keep track of subjects' emotions (see Bach, 2016, for a discussion of how it helps interpret data from economic experiments), and eye-tracking techniques for instance allow us to record how information is collected by participants in the course of the experiment (see Lahey and Oxley, 2016, for a review). Such measurement tools deliver 'control variables': empirical measures of the components of decision heterogeneity. In line with the identification strategy described in Section 3.4.2, these control variables narrow the scope of the unobserved variations leading to the outcome, thus enhancing the quality of inference. They can also be used to better interpret the results, by assessing the role played by such heterogeneity in the variations of interest in the outcome variable.

Illustration 5.10

Controlling for closeness: the inclusion-of-the-other-in-the-self scale

An important question in behavioural economics is the effect of pre-existing relationships between people – to what extent and why do members of the same family, friends, employees of the same organisation, people from the same country, etc., interact differently together. The usual way of studying this dimension is to ask people to report the nature of their relationship with others (or, similarly, to organise the experiment in such a way that people with a given pre-existing relationship come to participate). Such information provides a proxy variable of how people are linked together, but does not measures the actual strength of this link. To overcome this issue, Gächter et al. (2015b) develop a measure of intensity borrowed from social psychology, the 'inclusion-of-the-other-in-the-self' (IOS) task (Aron et al., 1992), based on the figure below.



For each possible choice, one circle refers to the respondent and one circle refers to another person, *X*. For a clearly designated participant, respondents are asked to 'consider which of these pairs of circles best describes your relationship with [this individual] in all questions that follow. In the figure "X" serves as a placeholder for [this individual], that is, you should think of "X" being [this individual]. By selecting the appropriate number please indicate to what extent you and [this individual] are connected'. Gächter et al. (2015b) confirm empirically the psychometric properties of the original scale, and show it describes not only 'close relationships' (typically, romantic ones) but also non-close relationships, in particular friends and acquaintances.

5.3.2 Treatment Parameters and Experimental Treatments

Examples abound in the experiments described in previous parts of the book of parameters that are purposefully set at different values. To name a few: the marginal per capita return in Isaac et al.'s (1984) experiment described in Section 5.1.4 is a treatment variable set equal to either 0.3 or 0.75; the alternative split against which subjects must choose in the discrete version of the UBG of Falk et al.'s (2003) experiment described in Illustration 5.8 is another example of *one* treatment variable, which takes *four possible values*: $(5,5),(2,8),(8,2)$ or $(10,0)$. Due to the frequent use of experiments in medical sciences, these kinds of parameter, which take more than one value according to the design of the experiment, are called *treatment parameters* (sometimes also called

treatment variables) – while parameters that remain constant in all instances of the experiment will be called *control parameters* (or control variables).

Treatment parameters need to be distinguished from *experimental treatments*: an experimental treatment usually refers to a unique combination of all the parameters' values. In the last example, one treatment variable defines four experimental treatments – the sender chooses between $(8, 2)$ and $(5, 5)$ in the first treatment, between $(8, 2)$ and $(2, 8)$ in the second, etc. In the first example, the two values for the MPCR define two experimental treatments.[5]

The main difference between control and treatment parameters is that treatment parameters aim to generate variations in the outcome, in order to provide identification of their causal effect. The causal effect of control parameters cannot be measured, precisely because they remain constant at a given value in all instances in which outcome behaviour is observed. But their effect can always be measured by implementing them as treatments rather than controls. As a matter of fact, the classification between controls and treatments belongs to the experimental design in just the same way as the value at which they are set: all examples of control parameters given above can be implemented as treatments, if measuring their effect on outcome behaviour is relevant or interesting. For this reason it is often very tempting to widen the set of treatment variables, and the number of values considered for each them. This temptation has to be resisted, due to the data requirements implied by such an inflation in treatment dimensions. The reason lies in the resulting rise in the number of experimental treatments necessary to provide identification – a question to which we now turn.

5.3.3 Between-Subject Designs: Identification of Treatment Effects through Randomisation

Consider an experiment with one treatment variable fixed at two possible values (as the MPCR example discussed above, for instance). This defines two versions of the experiment called experimental treatments (or sometimes experimental *conditions*):[6] one condition for each possible value of the treatment, with all other parameters fixed at a constant value. The open question we want to discuss in the reminder is: how should the two conditions be implemented to achieve identification of the effect of this treatment on outcome behaviour?

The discussions in Chapter 3, Section 3.4.1 in particular, provide a natural answer to this question. Identification is achieved if subjects are allocated to one condition or another according to a random rule. This produces two kinds of individual: those involved in the 'baseline', or 'control', condition and those involved in the 'treatment' condition. Denoting $\mathcal{T}_i = 0, 1$ the group variable indicating the condition individual i is involved in, the effect of the treatment on the outcome variable y_i can be estimated by simple mean comparisons:

[5] Treatment parameters and experimental treatments trivially coincide in experiments that consider two possible values for a unique treatment variable; the two wordings are often used as synonyms in such circumstances.

[6] In general, economists say 'treatment' where psychologists say 'condition'.

$$\widehat{\Delta} = \frac{1}{N_{\mathcal{T}_i=1}} \sum_{i \in \mathcal{T}_i=1} y_i - \frac{1}{N_{\mathcal{T}_i=0}} \sum_{i \in \mathcal{T}_i=0} y_i$$

Since the allocation to one group or another is random, the individual heterogeneity that makes y vary for unobserved reasons is not correlated to the experimental condition. This simple comparison thus provides an unbiased estimate of the mean causal effect of the treatment on behaviour. This identification strategy is identical to the cross-section estimator $\widehat{\Delta}^{Cross}$ introduced in Chapter 3. Randomisation of treatment groups achieves the identifying assumption of this estimator, i.e. that there is no selection into the treatment groups.

In the experimental literature, this implementation is often referred to as a 'between-subjects' design, since individuals involved in each condition are different people. People generally refrain from having different treatments implemented in the same session, in order to preserve common knowledge of equal treatment between subjects. As a result, between-subject experiments generally implement one condition per session. Identification requires the experimental treatment implemented in a given session to be chosen at 'random', i.e. in a way that is uncorrelated with any unobserved factor that might have an effect on behaviour. To get back to a previous example, if some sessions are scheduled early in the morning and others just after lunch, the mapping of these slots with the treatments should be chosen in such a way that there is no systematic relationship between the two. If it is the case, then people in the two groups are statistically the same: the distribution of noise in the treatment sub-samples is identical. Because the counterfactual is statistical in nature, between-subject designs are demanding in terms of sample size.

5.3.4 Within-Subject Designs: Identification of Treatment Effects through Blocking Strategies

The alternative to a statistical counterfactual based on randomisation is to get rid of the confounding effect of individual unobserved heterogeneity by just having the very same people behaving in all decision environments. In this way, unobservable individual heterogeneity would no longer interact with the treatment by remaining constant in all decision environments. This can be achieved by implementing experimental treatments one after the other in the same experimental session, with the same subjects. In the example of two experimental treatments, decision periods denoted $\underline{t}$ are associated with the first of them ($\mathcal{T}_{i,\underline{t}} = 0$, $\forall i$), and decision periods denoted $\bar{t}$ are associated with the second ($\mathcal{T}_{i,\bar{t}} = 1$, $\forall i$). The effect of the treatment can be measured by a simple comparison of outcomes between the two situations:

$$\widehat{\Delta} = \frac{1}{N} \sum_i y_{i,\bar{t}} - \frac{1}{N} \sum_i y_{i,\underline{t}}$$

Since the same individuals are observed in two different periods of time, this formally amounts to relying on the before–after estimator introduced in Chapter 3, Focus 3.4. It uses the past behaviour of treated subjects as a counterfactual for their behaviour

after the treatment. This discussion underlined that this estimator thus identifies a mix between the treatment effect of interest and the variation in unobserved heterogeneity in the group of treated individuals, from before ($\boldsymbol{\varepsilon}(1)_{t=\underline{t}}$) to after ($\boldsymbol{\varepsilon}(1)_{t=\bar{t}}$) the treatment. Since the same individuals are making decisions in the two situations, the difference $\boldsymbol{\varepsilon}(1)_{t=\bar{t}} - \boldsymbol{\varepsilon}(1)_{t=\underline{t}}$ eliminates any permanent heterogeneity. Individual-specific determinants of behaviour, in particular, disappear thanks to this comparison. Because the comparison occurs for the same individuals observed in different conditions, this design is often labelled a *within-subject* implementation of the treatment.

One advantage of this design is to achieve a higher statistical power than the between-subject design, thanks to both lower noise in the data and more individual observations delivered by a given sample of participants. First, within-subject designs generate observations associated with several different treatments based on the same individuals – i.e. it increases the number of subjects who participate in each treatment for a given size of the subject pool. Second, since individual-specific heterogeneity remains constant across decision stages, variations in the outcome variable are conditional on these unobservables, resulting in lower variance of the noise as compared to between-subject designs. Illustration 5.11 provides an example of how this conditioning on individual heterogeneity widens the sets of research questions that can be addressed.

The price for this increased power is that identification is weaker, as the identifying assumption is more likely to be violated. The identification assumption amounts to requiring that the treatment is the only influential change over time in the experiment. Any change in unobserved heterogeneity that happens at the same time as a new experimental treatment is introduced will thus be confounding. Typically, if repetition itself induces a change in decision-making, because subjects learn how to best decide, or get tired or bored with the experimental exercise, the estimator will confound this effect and the causal effect of behaving in environment $\underline{t}$ rather than $\bar{t}$. A major source of such a change over time is the sequence of the treatments themselves. If there is any permanent effect of being exposed to a treatment on subsequent behaviour, then the comparison no longer elicits the pure effect of behaving in one environment as compared to behaving in another, but rather a combination of the treatment effect of interest and this change in unobserved heterogeneity over the course of the experiment. Because such a confounding effect arises due to the order in which the treatments are implemented, it is known as *order effects* in the experimental literature (see Illustration 5.12 for an empirical example, in which order effects are induced on purpose).

As for any confounding mechanism, order effect can be indirectly controlled through either blocking or randomisation. Randomisation in this context implies choosing randomly the order in which treatments are implemented. There will be change over time in unobserved heterogeneity, but randomisation will ensure that this change is not systematically correlated with the occurrence of the treatments, hence achieving identification. As before, this identifying assumption is statistical in nature, hence requiring gathering enough data for the assumption to be empirically meaningful. Alternatively, a blocking design will systematically balance the order in which treatments are implemented.

Illustration 5.11

Individual consistency of social preferences: a within-subject design

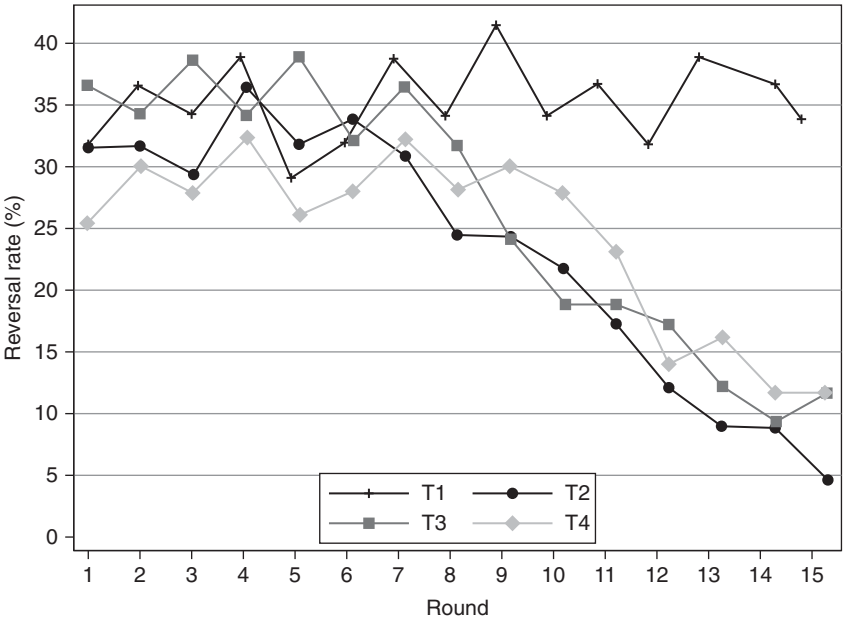
The inequality-aversion model of Fehr and Schmidt, presented in Section 5.2.4, rationalises behaviour in social-preference games by extending the specification of individual preferences to additional parameters, θ^-, θ^+, that relate individual well-being to the distribution of outcomes. The model has been shown to perform well in describing the observed distribution of decisions in social-preference games. Blanco et al. (2011) aim to assess whether such predictive power holds at the individual level. To that end, they design an experiment eliciting behaviour in social-preference games for the same individuals. This within-subject design allows us to measure the consistency of individual preferences across games. In this experiment, the same subjects plays the following games one after the other:

Game	Label	Description
Ultimatum-bargaining game	UBG	£20 pie, proposer gets £$(20 - s)$ and responder s if the respondent accepts, both get 0 otherwise
Modified dictator game	MDG	dictator chooses between £20 and £0 and equitable outcomes ranging from £0–£0 to £20–£20
Sequential-move prisoners' dilemma	SPD	both defect: £10–£10; both cooperate: £14–£14; one defects, one cooperates: £17–£17
Public-good game	PG	two players, £10 endowment per player, marginal per capita return on contributions is 0.7

From these data, individual preferences (i.e. estimates of θ_i^-, θ_i^+ for each subject i) are estimated based on behaviour as receiver in the UBG and as proposer in the DG. This provides a distribution of preferences in the sample of subjects. The analysis then proceeds in two steps. First, the distribution of preferences is used to assess its ability to predict aggregate behaviour: this amounts to comparing the actual distribution of preferences to the distribution required in order to generate the observed distribution of behaviour in each game. The model performs well in that regard in the ultimatum-bargaining game, the public-good game and the first-mover strategy of the sequential prisoners' dilemma, confirming previous analysis. The second step makes use of the within-subject dimension of the data. Individual decisions in each game are compared to the prediction from the individual-specific estimated preferences. Based on this within-subject consistency criterion, the model performance is very low. As an illustration, we focus on the results from the public-good game. People who exhibit a high level of guilt about advantageous inequalities ($\theta_i^+ > 0.3$) should make a positive contribution to the public good. On the other hand, people with a low guilt parameter are expected to free-ride on the public good. In the sample, 20 subjects are characterised by preferences such that $\theta_i^+ < 0.3$, and 17 subjects contributed zero, confirming the good performance of the model in aggregate. However, only 13 of the 20 subjects with low θ^+ belong to the group of zero contributors. The hypothesis that a low θ^+ implies free riding at the individual level thus is rejected by the data.

Illustration 5.12
Evidence of order effects: rationality spillovers

Cherry et al. (2003) rely on a within-subject design to test a 'rationality spillover hypothesis', i.e. that non-market behaviour changes when rationality is fostered through a market setting. The hypothesis is tested in the context of a preference-reversal game. The baseline, T1, features no arbitrage and real choices – i.e. choices affect take-home pay. Subjects play 15 rounds. In each of them, subjects choose between two lotteries (*A* and *B*) with the same expected value, first in a market setting and secondly in a non-market setting. At the beginning of each round, they are endowed with an initial balance of $10. In both settings, subjects are asked to order their preferences between the first lottery (*A*) – which is a low-risk lottery – and the second lottery (*B*) – which is a high-risk lottery. Then, they are asked to report their fair value for both lotteries in both settings. Finally, an offer price is randomly drawn for each lottery in the market setting and the subject is sold the lottery if the stated value was higher or equal to the price – the price is subtracted from the round balance – and the buyer becomes the owner of the lottery: the lottery is played, and earnings are determined according to a draw in the lottery. Arbitrage is introduced in T2, in order to test the rationality spillover hypothesis. Starting at round 6 in the market setting, preference reversals are automatically arbitraged: the lottery with the lowest price is sold to the subject, and traded against the preferred lottery. This realises the monetary cost of preference reversals. Two further treatments provide robustness checks. In T3, everything is similar to T2 except that in the non-market setting subjects make hypothetical choices instead of real choices. In T4, everything is similar to T3 except that subjects make a choice over environmental lotteries instead of monetary lotteries as in all other treatments. For example, a subject is asked to choose between seeing a grizzly bear with a 30% chance and catching a cut-throat trout with a 70% chance.



The main results from the experiment are provided in the figure above (from Cherry et al., 2003, p. 71, Figure 2). Thanks to the within-subject design, subjects serve as their own control in the non-market setting. Comparing the trends between T1 and T2, shows a strong discrepancy after 6 rounds (when arbitrage is introduced), in the number of observed preference reversals: the double difference confirms the rationality-spillovers hypothesis. This decrease in reversal rates is observed even when the choice is hypothetical (T3) and when the lotteries are environmental instead of monetary (T4).

In the example of two experimental conditions, this amounts to balancing the number of sessions between the two orders $\{\mathcal{T}_{i,\underline{t}} = 1; \mathcal{T}_{i,\bar{t}} = 0\}$ and $\{\mathcal{T}_{i,\underline{t}} = 0; \mathcal{T}_{i,\bar{t}} = 1\}$. The identifying assumption is more restrictive than the one associated with randomisation: it achieves identification only if the time-varying confounding mechanisms are the same in the two sequences – for instance: if it is not the case that one of the two conditions fosters quicker learning than the other. The design, however, extends the scope of the robustness check that can be performed in that regard. One can use only the data from the first half of each sequence to perform a between-subject analysis of the treatment effects. Cross-sequence comparisons can also be used to assess whether behaviour is different over time by comparing $\mathcal{T}_{i,\bar{t}} = 1$ to $\mathcal{T}_{i,\underline{t}} = 1$ and $\mathcal{T}_{i,\bar{t}} = 0$ to $\mathcal{T}_{i,\underline{t}} = 0$ (to distinguish order effects from pure learning effects, it might be necessary to consider another sequence, $\{\mathcal{T}_{i,\bar{t}} = 0; \mathcal{T}_{i,\underline{t}} = 0\}$ in order to produce the counterfactual behaviour in treatments $\mathcal{T}_{i,\underline{t}} = 0$ without order effects). Either of the two choices implies complementing the within-subject design by additional treatments with different orders. In the end, the number of experimental treatments, and thus participants, required for a within-subject design is thus the same as for a between-subject design.

5.3.5 Multiple Treatments

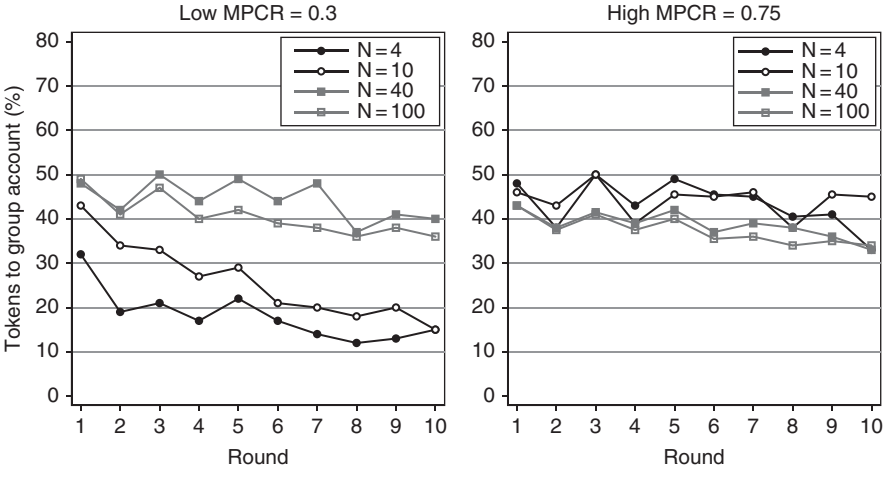
The distinction between treatment variables and experimental treatments is mainly relevant when there are more than two values of the treatment parameters. This happens when one treatment variable is set to more than two possible values, but more often so when several variables are used as treatments. Illustration 5.13 provides an example that combines the two cases – with two treatment variables in a VCM game, the MPCR and the group size, one of them associated with two possible values while the other can take either of four levels. To make things concrete, consider two treatment variables denoted $\mathcal{T}_a$ and $\mathcal{T}_b$, each associated with two values of the corresponding treatment variable, which we denote as usual as $\mathcal{T}_a = \{0, 1\}$ and $\mathcal{T}_b = \{0, 1\}$. The challenge is to define experimental treatments in such a way that the causal effect of each one can be identified from observed behaviour. Strict randomisation of the treatment variables would imply independently randomly choosing the parameter value of each treatment variable. The main drawback of this procedure is that it allows several treatment variables to change simultaneously.

The implementation rule circumventing this issue amounts to defining experimental treatments according to a *factorial design*: an experimental treatment is defined for

Illustration 5.13

VCM: a 4×2 factorial design

As shown in Focus 5.3, the intensity of the social dilemma raised by a VCM game crucially depends on how the MPCR compares to the individual shares in the group, $1/N$. Isaac et al. (1994) further investigate the empirical content of this prediction by considering variations in both dimensions. The core game is the one described in Section 5.1.4. Two dimensions are used as treatment variables: the MPCR is set at either 0.3 or 0.75, and the group size takes four possible values: 4, 10, 40 and 100. Together these two treatment variables define a 4×2 factorial design, resulting in eight experimental treatments. The figure below (from Isaac et al., 1994, p. 14, Figure 6) shows the share of contribution elicited over time in each treatment.



For a given private return to the public good, group size does not decrease subjects' contribution per se. In particular, for large groups, an increase in the group size has a positive influence on the contribution to the public good. Therefore, large groups can be more efficient at providing public goods than small groups. Two additional features appear from the figure. First, with a low marginal per capita return, large groups are more cooperative than small groups, but this effect disappears when the marginal per capita return increases to 0.75. Second, the positive link between the marginal per capita return and the contribution to the public good that existed in small groups vanishes in large groups.

each and every combination of all treatment variables. Table 5.4 illustrates the factorial design associated with the above example, with two treatment parameters each set to two possible values. This 2×2 factorial design results in four experimental treatments. The advantage of a factorial design is to build a control situation for each treatment condition: comparisons across cells in the table identify the marginal effect of switching the value of one treatment variable, conditional on some value for the other. It also delivers more: as shown by the comparisons in the outer row and column of Table 5.4, as many estimates of the treatment effects as the number of parameter values (two in our example) are observed thanks to this design. For instance, $\Delta_{\mathcal{T}_a|\mathcal{T}_b=0}$ and $\Delta_{\mathcal{T}_a|\mathcal{T}_b=1}$

Table 5.4 Multiple treatment variables: a 2 × 2 factorial design

		Treatment variable 2		
		$\mathcal{T}_b = 0$	$\mathcal{T}_b = 1$	$\neq$
Treatment variable 1	$\mathcal{T}_a = 0$	Experimental treatment 1	Experimental treatment 2	$\Delta_{\mathcal{T}_b \mid \mathcal{T}_a=0}$
	$\mathcal{T}_a = 1$	Experimental treatment 3	Experimental treatment 4	$\Delta_{\mathcal{T}_b \mid \mathcal{T}_a=1}$
	$\neq$	$\Delta_{\mathcal{T}_a \mid \mathcal{T}_b=0}$	$\Delta_{\mathcal{T}_a \mid \mathcal{T}_b=1}$	Δ^{DD}

both are measures of the effect of the treatment variable $\mathcal{T}_a$, but each is generated by a different decision environment in terms of the value of $\mathcal{T}_b$. The difference between the two provides a measure of the interaction between the treatments. It amounts to relying on the difference-in-difference estimator (DD), i.e.:

$$\begin{aligned}\widehat{\Delta}^{\mathrm{DD}} &= \left(\overline{Y}_{\mathcal{T}_a=1,\mathcal{T}_b=1} - \overline{Y}_{\mathcal{T}_a=1,\mathcal{T}_b=0}\right) - \left(\overline{Y}_{\mathcal{T}_a=0,\mathcal{T}_b=1} - \overline{Y}_{\mathcal{T}_a=0,\mathcal{T}_b=0}\right) \\ &= \left(\overline{Y}_{\mathcal{T}_a=1,\mathcal{T}_b=1} - \overline{Y}_{\mathcal{T}_a=0,\mathcal{T}_b=1}\right) - \left(\overline{Y}_{\mathcal{T}_a=1,\mathcal{T}_b=0} - \overline{Y}_{\mathcal{T}_a=0,\mathcal{T}_b=0}\right)\end{aligned}$$

When applied to a design in which the treatments all are target variables, this estimator measures the joint contribution of the two treatments, i.e. how the outcome varies when the two treatments are simultaneously influential, as compared to their own marginal effect. This same strategy is sometimes applied to designs in which one of the two treatment variables is used to generate the baseline: the target outcome, on which the treatment effect is to be identified, is the change in behaviour between two baseline conditions. It is the case, for instance, in experiments described in Section 2.4, trying to measure the hypothetical bias in preference elicitation. In this case, the variation in behaviour depending on whether incentives are real or hypothetical (say, treatment variable $\mathcal{T}_a$) serves as a benchmark for the investigation. The main treatment variable of interest is another dimension of the environment ($\mathcal{T}_b$, such as a priming task, or a certainty question) of which the design aims to measure the effect on hypothetical bias in elicited preferences. In circumstances like this, the difference-in-difference estimation strategy is applied in order to measure the effect of the treatment variable of interest ($\mathcal{T}_b$) on the variation of behaviour. As shown in Chapter 3, Focus 3.4, the identifying assumption is then the so-called parallel-trend assumption: that the change in unobserved heterogeneity that occurs in line with the 'nuisance' treatment variable is not affected by the change in the target one.

The factorial design implementation of multiple-treatments experiments is the main reason why the choice of treatment variables, and their individual values, must be made with parsimony. In the general case, with K treatment variables with n_k ($k = 1, \ldots K$) individual values each, a factorial implementation requires as much as $\prod_{k=1}^{K} n_k$ experimental treatments, defining the set of all feasible combinations of the treatment variables. For a given number of observations, this implies an important loss of statistical power as additional treatment dimensions are added; conversely, for a given statistical

power, the required sample size explodes as the number of dimensions increases. As an example, adding one value to one of the two treatment variables in a 2×2 design increases the number of treatments from 4 to 6 (2×3), and adding a third binary variable increases the number of conditions from 4 to 8 ($2 \times 2 \times 2$).

5.4 The Perceived Experiment

The quality of identification in an experiment is based on two building blocks. The first is a well-designed decision environment, as discussed up to now. But the actual decision environment is not quite the experiment that has been designed and implemented. What actually generates outcome behaviour is rather the experiment in which participants *think* they are involved. This perceived experiment is the second building block of internal validity. For instance, if subjects believe they will not get paid at the end of the experiment, then the performance-based reward will not incentivise behaviour; if they think the worst outcome is always to be drawn to manipulate their payoffs, then the experiment will not deliver their actual decisions under risk, etc. No matter how clever the actual design is, outcome behaviour will only result from what people have in mind – see Illustration 5.14 for an application to the incentive-compatibility of the compensation scheme. As a result, control over subjects' perceived experiment is a very important part of experimental designs, and their internal validity.

5.4.1 The Perceived Situation: Experimental Instructions

Within the set of controls over the perceived experiment, the leading one is the way the experiment is explained to the subjects. This is usually referred to as the '*experimental instructions*', and generally takes the form of a printed sheet of paper that is distributed to subjects before the experiment starts. Section 2.1 provided a concrete example applied to a second-price Vickrey auction; Section 6.2.1 provides practical advice about its structure and content. What matters in terms of internal validity is that this document is the main source of information for subjects about how the experiment proceeds. As such, it has to be written in such a way as to fulfil two basic aims: (i) that the experiment is well understood by each and every subject, and (ii) that all subjects are given the same information about the experiment. Each of the two conditions has its own consequences on the internal validity of the outcome behaviour. Condition (i) guarantees that the decision environment that subjects have in mind is the one that ought to be implemented. Otherwise, observed behaviour will either be generated by another experiment (the one subjects think they behave in) or, even worse, by random behaviour in an environment subjects simply did not get. Condition (ii) applies a blocking strategy. The aim is to avoid idiosyncratic noise in the perceived experiment that would be induced by a heterogeneous understanding of the rules – hence preventing correlation with any change in target outcomes. Each condition also has its own implications on how experimental instructions are written and communicated to subjects.

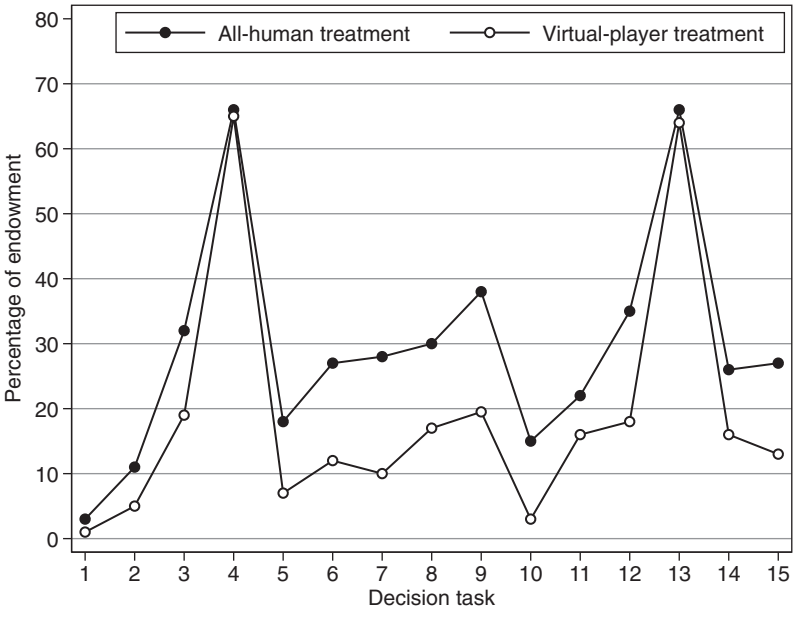
Illustration 5.14

Identified failures of internal validity: confusion in VCM games

Ferraro and Vossler (2010) further explore Andreoni's (1995) hypothesis presented in Section 5.1.4 by considering whether players' inability to distinguish the relationship between their choice and the game's incentives plays a role in cooperation behaviour in VCM games – a feature they label '*confusion*'. They design a series of experiments which compare VCM outcomes when the game is played with other humans or with virtual players. Virtual players are designed to perform predetermined contribution sequences. Subjects are informed whether they are matched with real people or with robots. In a virtual-player treatment, other-regarding preferences should play no role.

To test this hypothesis, participants are involved in 15 VCM games, with varying MPCR and group sizes – in order to check the robustness of the difference in changes in the incentives. The nature of other players (whether they are automatons or real human beings), by contrast, remains the same in all decision rounds. The main results are shown on the figure (from Ferraro and Vossler, 2010, p. 9, Figure 1).

Based on the comparison between the two treatments, confusion accounts for half the average contribution – equal to 25% in the human treatments, 12.5% in the virtual-player treatments. What is more, confusion does not seem to disappear with repetition.



The second condition, requiring that the same information is given to all subjects, is the main reason why experimental instructions are actually written and distributed to subjects. Thanks to this document, explaining the experimental procedures to subjects amounts to reading the same text to all subjects: the same wording, the same order, the same sentences, the same examples will thus be used to communicate the decision

environment to subjects, avoiding variations between subjects across experimental sessions. For this reason, it is also very important to refrain from giving information on an idiosyncratic basis. In particular, it is always very tempting to use the written text as a simple memo and explain the experimental set-up using one's own words as they come: but doing so incurs the risk of forgetting the exact wording, or even pieces of information, used in past sessions and subsequently inducing variations in perceptions across subjects. Similarly, it is good practice to always try to refer to written instructions when answering questions. If the instructions are well written, all information subjects need to know should be described; if they are not, the right answer to a question about it should simply be: 'you don't have this information; it's not specified'.

The first condition requires the decision environment to be well understood by each and every participant. The obvious consequence is that the instruction must be simple, intelligible and clear. This has to be true for every subject possibly drawn from the pool. Smart people easily understand simple things, but the reverse is not true. The instructions thus have to be as accessible as possible, even if some subjects are very likely to find the text overly simple, and the explanations very boring – no doubt, this is what you felt while reading the sample instructions in Chapter 2, Section 2.1. Our advice when writing the experimental instructions is to try to figure out what the least talented reader you are able to imagine would think based on the text. This is not only a matter of IQ, but also very often a matter of jargon and intuitions that are obvious to people well versed in economics, but not quite so to others. Here are a few principles that are generally applied to the writing of the instructions:

- Only focus on the information subjects need to know – many features of the design are not necessary to understand decisions to be made.
- Describe, do not explain – the reasons why things work the way they work is a matter of inference, explaining them will warn subjects against particular kinds of behaviour.
- Never repeat the same information – otherwise, subjects will wonder why you are giving the same information several times, and will conclude that they misunderstood the first time they heard the information.
- Smooth out any possible ambiguity – each time subjects wonder about some feature of the design, they will make out their own answer if it is not immediately provided.
- Parsimoniously choose the place to provide each piece of information – answers to questions one does not have in mind are generally skipped.
- Always use the same words to refer to the same things – people generally expect different meanings for different words.

Beyond writing, two devices are useful and very often included in order to enhance the level of understanding of the experiment. First, it is advised to include many examples illustrating the relationships between the information received, the decision taken by one subject and the others, and the resulting outcomes. General ideas always are more easy to get once their concrete applications have been presented. There is a risk of anchoring subsequent behaviour through the examples, though: typically, the likelihood of observing high contributions in VCM game will be higher if the examples show high

levels of contributions (because the examples are interpreted as social norms, or advice from the experimenter, see e.g. Roux and Thöni, 2015, for empirical evidence). Several strategies can be used to circumvent this issue: provide several examples, in random order, that are balanced as regards the kind of behaviour they are likely to induce (high, medium and low contributions, for instance), or use letters instead of numbers (as was the case in the instructions provided in Chapter 2). One can even ask subjects to make up their own examples, which implies the ability to check subjects' answers one by one. In any case, the choice of the examples must be consistent with the level of cognition expected from the subjects. If the relationships between decisions and outcomes require some computation, several possibilities exist. One can offer only a few examples along with a computation formula, or provide a table summarising all possible configurations, or a calculator/simulator which subjects are allowed to use freely during the experiment. Each choice, as always, has its own advantages and drawbacks. The less easy it is for subjects to compute the consequences of their choice, the more likely are computational mistakes; conversely, the more sophisticated are computational tools, the more likely it is that the design artificially forces payoff maximisation. Requate and Waichman (2011) provide empirical evidence for the case of a duopoly game, proving that this choice is not innocuous on elicited behaviour.

A second device is the use of pre-experiment understanding questionnaires, which subjects are typically asked to answer on their own after the reading of the instructions and are then publicly commented on before the experiment starts (an example of both the questionnaire and how the debriefing takes place has been provided in Section 2.1). The same rule applies to the debriefing as the ones stated about the written instructions: to avoid variations in the perceived environment across experimental sessions, it is wise to debrief the questionnaire using a standardised wording as close as possible to the written instructions. This questionnaire allows us to double check what subjects got from the instructions, and provides an opportunity to fix possible misunderstandings. This is also a good place to summarise the main important information subjects should keep in mind from the instructions. In that regard, even again if the questions might seem overly simple to most participants, it is always a good idea to include questions about the occurrence of the main treatment variables and the most sensitive features of the design.

The above guidelines are all meant to achieve the highest possible level of understanding about what the actual experiment is. A related but different concern is to try to undermine the heterogeneity of perceptions induced by the psychological load each participant associates with the decision environment. This has two main consequences for the way instructions are devised in typical economic experiments. First, instructions are generally written in 'neutral' words, i.e. do not make reference to either the actual economic context of the situation, or to the actual name or role of the economic agents involved. This is meant to focus on economic incentives and avoid eliciting the mechanical effect of norms or value judgements that subjects project onto the underlying real-world situation. For instance, a game on lying behaviour will not make use of words like 'telling the truth' or 'lying', but rather describe several 'options' and their consequences. Similarly, an experiment focusing on corruption behaviour will not talk

about bribes between a briber and a bribee, but rather describe it as monetary transfers between player A and player B. The usual disclaimer applies to this discussion: it applies to experiments in which moral judgements induced by the situation are confounding, but in other instances they might well be a matter of interest on its own, leading to purposefully contextualised instructions. Empirical evidence in that regard is inconclusive and game-dependent: for instance, Abbink and Hennig-Schmidt (2006) find no effect of the wording on corruption behaviour, while Jacquemet et al. (2017) find a strong effect on lying behaviour. A strong limitation of these principles is that contextualised wording often eases understanding (see Illustration 5.15 for an example). For instance, explaining a market experiment without words like price, supply, demand, good, seller and buyer sounds like an impossible challenge, resulting in a uselessly complicated text. Experimental evidence tends to confirm this point, as a meaningful context has been shown to work as a substitute to learning (Cooper and Kagel, 2003, 2009; see Alekseev et al., 2017, for a literature review).

5.4.2 Never Use Deception

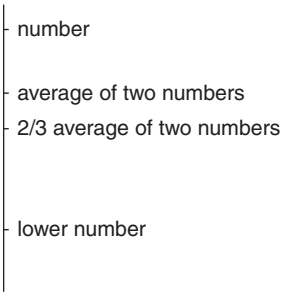
An implicit but crucial assumption in the implementation of the experiment is that subjects trust what they are told. All the above rules are nothing but useless if subjects doubt the information they receive. If that is the case, then what they actually understand about the experiment proceedings will be a mix of this information and what they think is the truth behind the fake details they are provided with. Such a risk of facing scepticism about the information provided in the instructions is greatly fostered by the use of deception, i.e. if false information is deliberately given to subjects. There are many channels through which subjects might realise that deception has been used in the experiment in which they have been involved. If deception applies to draws being random, they can talk together after the experiment and then observe that highly unlikely outcomes occurred (for instance, that the low-stake outcome has always been drawn although a payoff-improving one was announced to be very likely); they can read the research papers derived from the experiment, and observe that the actual design is different from the one they were described at the time the experiment was run, etc. The most important and devastating consequence of deception is that it is associated with the possibility of contamination effects: if deception is used for one experiment in one laboratory, this might induce mistrust not only from subjects from the same pool who participated in the experiment, but even from any subject from the same pool or even in any other laboratory. The consequence is that control over the perceived environment is then lost, as there is no way to know for sure whether subjects actually take the instructions at face value or think they were deceived at some point (see Ortmann and Hertwig, 2002, for a review of empirical evidence).

This feature is obviously shared by all experimental social sciences. In economics (in contrast, in particular, with psychology: see Focus 5.6 for a discussion) such a practice is generally not necessary to produce interesting data: the focus is on behaviour generated by an environment that is most often easy to actually implement and describe. There is thus no need to manipulate the environment through false information. Given the small

Illustration 5.15
Identified failures of internal validity: game form recognition in beauty contest games

In a guessing game, each player in a group of N is asked to choose a number between 0 and 100. The winner is the player choosing the number closest to a given percentage p of the average of all numbers chosen in the group, $\overline{y}$. If $N \geq 2$ and $0 \leq p < 1$, the only Nash equilibrium is to choose 0. Accumulated evidence (which we review in more detail in Chapter 9, Section 9.2), however, shows that people very often choose numbers that are higher than 0 – a weakly dominated strategy. Chou et al. (2009) design an experiment aimed at assessing whether such behaviour should be seen as an actual failure of the game-theoretic prediction, or may be due to an internal-validity issue, noting that '*if subjects do not understand the game form, the experimental control needed for testing game theory is lost*' (Chou et al., 2009, p. 159). The experiment includes three treatment variables that are expected to improve subjects' understanding of the game. First, three different types of instruction are used to explain the game: the standard instructions are used in the baseline, while the HINT condition includes the following addition:

Notice how simple this is: the lower number will always win (see Figure)



The figure offers an explicit illustration of how one's own choice interacts with the choice of others. This aims to help subject figure out how the strategic interaction works. The second treatment variable is the scenario used in the instructions. In the BATTLE condition, the game is framed using a battle situation: subjects are told that they have to decide where to place their troops on a hill. Within each pair of players, the one with troops at the highest height will win the battle, and the hill is 100 feet high. Last, the experiment uses two kinds of sample, with varying levels of training in maths: Caltech and community college students. The main results are described in the table below (from Chou et al., 2009, p. 170, Table 4).

Measure	Caltech	Caltech HINT	College	College HINT	College BATTLE
Share with number = 0	0.46	0.87	0	0.07	0.46
Average number	23	3	35	31	15
Sample size	26	23	20	15	105

For all groups, the hint greatly helps elicit the equilibrium action. The change is stronger for Caltech than for community college students, who strongly benefit from the BATTLE framing. To the extent the treatment variables actually improve the clarity of the instructions, an enhanced understanding of the game seems to foster equilibrium play, hence challenging the accuracy of previously observed evidence.

Focus 5.6

Economics and psychology: an overview of the main methodological disagreements

Economics and psychology have a lot in common: they both aim to better understand individual and social behaviour, and both use controlled experiments to do so. Despite this strong proximity, their methodological standards evolved in different directions. This might explain why the two disciplines have long experienced difficulties talking to each other (Ariely and Norton, 2007), and also recovers some differences in the general approach and focus like, typically, the use of formal theoretical models in economics that are absent from psychology (Camerer, 1996). We provide below a list of the most consequential differences in the way experiments are run in both disciplines (those that will most likely elicit scepticism from a reader versed on the other side). Most of them have been raised in the seminal discussion provided by Loewenstein (1999), and are part of the critical reviews that can be found in, e.g., Hertwig and Ortmann (2001); Croson (2006); Ortmann (2010); Madsen and Stenheim (2015) along with empirical evidence from the two disciplines. Since this chapter has extensively discussed the usual practice and rationale for each dimension from the point of view of economic analysis, we focus below on the current practice in psychology.

- *Financial incentives*. Experiments in psychology generally do not involve performance-based incentives – while a fixed fee is sometimes offered to incentivise participation and compensate the opportunity cost of time. This difference lies in the driving forces of behaviour each discipline aims to study: in psychology, intrinsic motivation and psychological factors are the main focus, to which external incentives only add noise and unwarranted variations.
- *Context*. In psychology, the context in which decisions are made are often part of the dimensions of interest, while economics relies on theories that aim to be context-free. Experiments in psychology thus generally explicitly refer to this context.
- *Subject pool representativeness*. The above difference in approaches also has consequences on the subject pool. The characteristics of the subject pool are just as important as the decision-making environment in psychology, in such a way that the pool must be chosen in accordance with the scope to which the theory applies. It does not mean that representativeness is of no importance in economics. But this question is generally seen as a matter of external validity rather than internal validity (see Section 8.3.3, Chapter 8).
- *Repetition*. The usual practice of repeating several times the same decision task is highly specific to economics, and seems unacceptably artificial to most psychologists. Again, the scope of theories seems to be the main reason for this discrepancy: initial responses in a highly abstract environment from fresh decision-makers is of little interest in the aim of testing rational-choice theory; while spontaneous behaviour in a well-described real-world situation is the obvious outcome of interest in psychology.
- *Deception*. This is both the most controversial and the most systematic difference between the two experimental approaches, as discussed in the text.

returns to be expected from deception, and the high risk of failure in the internal validity it incurs, the use of deception is usually banished from experiments in economics. This principle stands as one exception to the usual disclaimer repeated many times in this chapter: there is unanimous consensus among scholars to never, in any case and for any

reason, lie in the instructions on the actual rules of the game (see Bonetti, 1998, for a review and a discussion of the cost–benefit analysis of this practice).

While the general principle, and its rationale, are both simple, its application is sometimes complicated. First of all, avoiding deception does not mean giving complete information. In many instances, there is no risk of fostering future mistrust by hiding some information from subjects. Typically, there is no need to know what exact matching mechanism is implemented in order to group people in the laboratory – remaining silent on this feature is not deceiving subjects, just efficiently selecting the information. More subtle and debated is the question of when deception starts if there is a discrepancy between the information given to subjects and what is actually implemented. As a typical example, consider an experiment in which a series of parameter values is chosen randomly and the distribution is announced to subjects (e.g., 10 values are independently drawn from a Bernoulli distribution with probability p). If the actual distribution is different from this one (the probability is p' or draws are serially correlated) then it is clearly deception. But what if the sequence is drawn from this distribution, but then the same sequence of draws is used in all instances of the experiment? There is quite strong disagreement among academics on the question. Whether or not this kind of practice is indeed deception is a matter of judgement, on which we thus refrain from giving any opinion. But in the end the criterion is whether or not subjects who would face the true information would feel cheated as compared to the actual proceedings they have been described.

5.5 Perceived Opponents and Learning

The previous discussion about deception shows the importance of beliefs about the design as a driving force of individual behaviour – and hence as a matter of concern in terms of internal validity. Another aspect on which participants will have to form beliefs is the other people involved in the experiment. This mainly concerns two dimensions: out-of-the-lab reputation refers to perceptions about the opponent that are inherited from outside the laboratory, while in-the-lab reputation will refer to changes in perception that occur in the course of the experiment. We will consider each in turn. In both cases, it is worth noting that such changes in beliefs will not be confounding if they are observed and can thus be used as conditioning variables. Belief-elicitation methods will be the topic of this chapter's case study, in Section 5.6. These methods are now well-established and useful in many circumstances. Whether they are enough to solve the internal-validity issue raised by changes in individual beliefs still is an open question, though, as it is not clear how belief elicitation and behaviour elicitation interact – see Illustration 5.16 for an example.

5.5.1 Out-of-the-Lab Reputation

In virtually all interaction contexts, people are likely to behave differently depending on what they know about others. Better knowing others helps anticipate what they will do in

Illustration 5.16

Belief elicitation and outcome behaviour in a VCM game

Croson (2000) investigates whether belief elicitation affects behaviour in a VCM experiment. Twenty-four subjects are arranged in six groups of four and kept in the same group across the games. Subjects are asked to play two games of ten periods each, and are informed after each period of the total contribution and of their own earnings. Each subject is endowed with 25 tokens, which could be allocated either to the private good, with a return of two cents per token, or to the public good, with a return of one cent per token to each member of the group. In the treatment group, subjects are involved in an additional step during which they are asked to estimate the contribution to the public good of the other three members of the group. Subjects are rewarded according to the precision of their predictions, up to 50 cents for an exact guess. The table below (from Table 1 in Croson, 2000, p. 305) shows the average contributions by group (control versus guess treatment). The top part of the table displays behaviour in the first set of 10 periods, the bottom part refers to the second set. Differences that are significant at the 5% level appear in bold.

	Round									
	1	2	3	4	5	6	7	8	9	10
Control 1	13.96	12.83	11.42	12.33	12.33	11.88	9.92	7.79	9.04	4.54
Guess 1	8.92	7.25	7.71	7.25	6.00	6.79	4.08	3.58	1.96	0.54
$\neq$	**5.04**	**5.58**	**3.71**	**5.08**	**6.33**	**5.08**	**5.83**	**4.21**	**7.08**	**4.00**
Control 2	11.54	11.33	10.29	7.88	7.33	6.88	4.21	6.50	4.25	2.67
Guess 2	3.79	2.29	1.58	1.21	1.71	2.63	2.54	2.67	2.50	1.96
$\neq$	**7.75**	**9.04**	**8.71**	**6.67**	**5.63**	**4.25**	**1.67**	**3.83**	**1.75**	0.71

As compared to behaviour elicited in the baseline, contributions in the treatment group are significantly lower. Full free riding is also observed to be more frequent, even if it is not enough to explain the lower average contribution. The evolution between games also seems to be affected: the restart value in the treatment group is about half the initial response in the first game, while the decrease amounts to less than 20% in the control group. Overall, having subjects think deeply about the behaviour of others seems to foster equilibrium play.

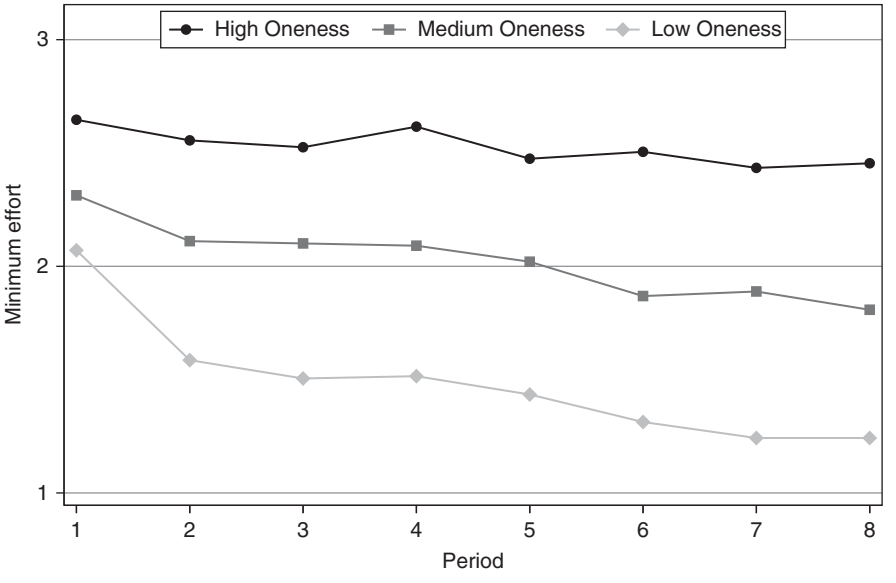
a given context, what kind of outcome they will favour, what behaviour they expect from others and maybe how kind one wants to be with them – the evidence provided in Illustration 5.17 confirms this case within the context of an experiment in which out-of-the lab relationships are a target treatment variable, rather than part of the noise. For all these reasons, unwarranted relationships between subjects can challenge internal validity. The general principle derived from this concern is to try to minimise uncontrolled variations through the design of the experiment.

This first implies choosing the subject pool in such a way that people are unlikely to know each other outside the laboratory. This is achieved by having participants come from various occupations, with different grade studies, universities or firms. Once

Illustration 5.17
The effect of closeness and the ability to coordinate

A coordination game captures the idea that value can be created when people coordinate their non-cooperative actions in a strategic environment (see e.g. Schelling, 1960; Cooper et al., 1990). 'Coordination failure' arises when people fail to attain the best outcome. This is typically due to strategic uncertainty – the uncertainty associated with not knowing how your opponent will play the game (see e.g. the survey by Devetag and Ortmann, 2007). Gächter et al. (2015a) aim to assess whether the pre-existing relationships between players of a coordination game helps overcome coordination failures. To that end, participants are invited to come to the experiment with three close acquaintances. When they arrive at the laboratory, they are part of either an F-MATCHING treatment – all four players are matched together – or an N-MATCHING treatment – participants are split into different groups. This generates exogenous variations in the pre-existing relationships in different groups. This is captured by the main treatment variable of interest, measuring the intensity of the relationships based on the inclusion-of-the-other-in-the-self scale presented in Illustration 5.10. The coordination game is a 'minimum-effort game' (aka, weak-link game, first introduced by Van Huyck et al., 1990) in which all four players need to decide simultaneously on a costly level of effort. The payoff everyone in the group earns is solely determined by the minimum of all group members' effort.

The figure (from Figure 5 in Gächter et al., 2015a, p. 18) shows the distributions of minimum efforts (on the y-axis, chosen by design between 0 and 5) across all periods of play (using a partner-matching design) split according to three categories of closeness. The ability to coordinate increases thanks to the strength of pre-existing relationships, with a higher initial level of cooperation as well as a steeper decrease over time.



arrived, people will have to wait until everybody is arrived. The waiting room/space must be designed in such a way that as little communication as possible occurs between participants. In particular, it requires the physical presence of the experimenter, or an assistant, in order to answer questions and provide information. These safety procedures will not prevent people from having heterogeneous beliefs induced by others' appearance, age, skin colour, etc. On top of that, anonymity is further maintained inside the laboratory so as to minimise the influence of any pre-existing partner-specific belief. This is achieved by having individual computers settled in cubicles equipped with separation walls between computers. This is complemented by trying to avoid by all means the identification of other participants in the course of the experiment. Computers will be matched randomly in the room, rather than using a physical criterion based on computers' positions in the room. Information transmission will be synchronised at the session level, rather than on the go, to avoid identification based on decision-making. Similarly, if roles are asymmetric and the number of decisions to be made is not the same, then the number of times a participant is using the mouse might be a cue about what role these subjects hold in the group. The interface should thus be designed in such a way that the number of clicks is the same in all roles – by adding requests to confirm one's choice, for instance.

5.5.2 Repeated Interactions: Partner-versus-Stranger Design

Out-of-the-lab reputation is the only concern about a player's belief regarding their partner if decisions in the experiment are one-shot – if only one decision is elicited. Repeated interactions are, however, often required, for two broad kinds of reasons – each associated with a different matching scheme between players, inducing varying statistical properties of the data.

The first obvious reason is to replicate a repeated game. In this case, the same players will interact together several times knowing that the composition of the group remains the same. Because this implementation purposefully allows players to build in-the-lab reputation (decisions at one stage depend not only on its current consequences, but also on its effect on future outcomes), it is known as a *partner design*. From a statistical point of view, the data produced by such a design are by construction serially correlated at the group level – outcomes at a given stage in the repeated game are influenced by decisions previously taken by both players. In terms of behaviour, variations over time result from both learning at the individual level and reputation building at the group level.

In some instances, the two need to be disentangled. It is the case when the purpose of the experiment is to replicate a one-shot game, while at the same time allowing for individual learning to enhance understanding of the game, and avoid eliciting initial responses to the environment. In this case, several repetitions of the same game are used as a means to allow subjects to become 'familiar' with the game and its rules. Such repeated versions of one-shot games are seen as a way to address the critique of Wallis and Friedman (1942, pp. 179–180, cited in Roth, 1993):

> It is questionable whether a subject in so artificial an experimental situation could know what he would make in an economic situation; not knowing, it is almost inevitable that he would, in

entire good faith, systematise his answers in such a way as to produce plausible but spurious results. For a satisfactory experiment it is essential that the subject give actual reactions to actual stimuli . . . Questionnaire or other devices based on conjectural responses to hypothetical stimuli do not satisfy this requirement. The responses are valueless because the subject cannot know how he would react. It allows for enough trials and errors for the data to be a convincing test of theoretical predictions (Binmore 1999).

A (perfect-)*stranger design* can be used to achieve both goals at the same time: it amounts to having subjects play the same game several times, but each time with a different partner. In practice, members of the group are randomly rematched at the beginning of each round. This is common knowledge among subjects, in such a way that there is no incentive to invest in reputation, despite the repetition of the game. This scheme strongly constrains the number of repetitions that can be implemented – with two subjects in a pair, the number of repetitions cannot exceed half the size of the session. To weaken the constraint, it is possible to implement a *pseudo-stranger design* in which participants are randomly rematched at the beginning of each round, but the probability of meeting the same others in the future is higher than 0. Because of this feature, there is an incentive to invest in reputation, albeit it is as small as the probability of meeting the same person again in subsequent occurrences of the game. The pseudo-stranger design is less robust to reputation building, but relaxes the constraint on the number of periods that can be played. From a statistical point of view, both kinds of stranger design are quite demanding in terms of data: because all players from a given session happen to play together at some stage in the experiment, the data are correlated at the session level. In contrast with the partner design, such correlation occurs not only within groups, but also across groups due to rematching over the course of the experiment. This drastically reduces the number of independent observations.

No matter what the matching design is, the number of repetitions itself should be chosen with parsimony. If enough learning is necessary to avoid measuring responses from confused subjects, too much repetition might induce subjects to seek what they think is the expected behaviour. As stressed by Camerer (1996), subjects might at some point be projected in the same position as Bill Murray in the famous 1990s movie *Groundhog Day*, in which the character adapts behaviour while living the same day again and again in order to eventually have Andie MacDowell fall in love with him. This is an example of an experimenter-demand effect, a challenge to external validity that will be discussed in Chapter 8.

5.6 *Case Study*: Eliciting Beliefs

Participants' beliefs in an experiment can change their behaviour, and prevent the experimenter from drawing firm conclusions about participants' motivations. For example, an even division of the endowment by the proposer in an ultimatum-bargaining game is compatible with fairness but also with the strict maximisation of private utility with well-specified beliefs (Manski, 2002). In addition, we saw that the perception of participants regarding the experiment itself, such as its aim, can also change behaviour. In this

case study, we consider whether beliefs can be measured in the lab. For the interested reader, more detailed surveys on methods for eliciting beliefs appear in Schlag et al. (2013) and Schotter and Trevino (2014).

The objective of belief elicitation is to uncover the likely value of some unknown event. An attractive representation of these beliefs is the probability distribution for the unknown event, taken to be a random variable. This is compelling as it allows the use of probability theory for the measurement of beliefs, offering a well-defined numerical scale for experimental answers. The validity of this representation of beliefs is usually assessed using two distinct criteria. First, the algebra of probability can be used to examine the internal consistency of beliefs. For example, many experiments evaluate the additivity of elicited probabilities. Second, beliefs can be directly matched with real frequencies to check whether beliefs are well or badly calibrated. The experimental-economics methods used to measure beliefs are usually imported from decision analysis, where they were originally designed to elicit expert opinions. As the focus was on experts, calibration was much more a concern than was the internal consistency of answers. We begin this case study with an overview of the current methods used to elicit beliefs in experiments. We then turn to the interaction between belief elicitation and behaviour in experiments.

In what follows, uncertainty is modelled via a state space $\mathcal{S}$. Only one of the states will come about, but the decision-maker does not know which one. Events E are subsets of $\mathcal{S}$. The complementary event of E is denoted E^c. The belief that event E will occur is denoted $\mathcal{B}(E)$.

5.6.1 Elicitation Methods

A number of methods have been proposed in the literature to elicit beliefs. One of the simplest is to ask participants directly about their beliefs. This method has the advantage of simplicity but is not based on individual choices. Regarding the analysis of economic behaviour, economists prefer to base their evaluation on revealed preference, i.e. the choices made by individuals. Choice-based elicitation procedures, such as scoring rules or matching probabilities, are then often preferred in experiments to introspective judgements. Focus 5.2 shows how prices can be used to predict unknown events with prediction markets.

Introspective Judgements

Asking about introspective judgements is the easiest and fastest way to elicit beliefs. The method is easily explained to subjects and corresponds to the common practice of eliciting expectations for binary events in surveys – see Illustration 5.18 for a discussion of their informativeness. A simple question to elicit beliefs about an unknown event via introspective judgement is the following:

What do you think is the percent chance that event E will occur?

Typical examples of event E in experiments are the strategy of another player in an experimental game, a price increase (or decrease) in a market, the identity of opponents,

Illustration 5.18

The accuracy of self-reported expectation measures

In a comprehensive survey recommending the increased use of expectation measures to understand choice behaviour, Manski (2004, Section 6) reviews the methods used to assess the accuracy of self-reported expectations about future events elicited from survey respondents. The aim of these assessments is to measure the validity of the rational-expectations framework: Are objective events accurately anticipated in subjective probabilities? It also provides evidence on how informative self-reported expectations are about the perceived likelihood of real-life events.

- When time-series data are available, the most obvious check is to compare the distribution of expectations to the distribution of outcomes in the same sub-sample.
- Using repeated cross-sections, if the cross-sections are randomly drawn from the same population (and there is no serial correlation in the outcomes across respondents), one can compare the distribution of outcomes in $t+1$ for current respondents to the expectations in t from a different sample.
- Last, this same assumption applies not only to forward-looking distributions of events, but also to past values: statistical independence between samples implies that past distributions of events can be used to assess the reliability of current expectations.

To illustrate the typical types of results from these methods, the table below (from Table 5 in Dominitz and Manski, 1997, p. 1362) compares the expectation over the next 12 months of 1994 respondents regarding (i) their chance of being covered by health insurance, (ii) the likelihood of experiencing a burglary, and (iii) the risk of losing their job; to the outcome distribution for these events among the 1995 respondents in the same survey.

	No health insurance		Victim of burglary		Job loss	
	Exp	Real	Exp	Real	Exp	Real
Male	0.15	0.15	0.16	0.05	0.15	0.18
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)
Female	0.16	0.13	0.17	0.03	0.21	0.18
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)

The table reveals a close match between mean expectations and outcomes for both health insurance and job loss, for both men and women. On the contrary, respondents tend to overestimate the likelihood of being a victim of burglary.

etc. Experiments typically ask subjects to evaluate their beliefs in percentage terms by selecting a number on a scale between 0 and 100. Rather than asking subjects to directly report percentages, verbal expressions of likelihood can be used, along with numerical probability categories that the event will occur. Manski (2004) recommends using a range of probabilities to reflect that subjects might have limited confidence in their beliefs. Illustration 5.18 describes methods that can be used to assess the accuracy of self-reported expectations about future events. In this case, subjects can express not only various degrees of belief (such as 100% for certainty, 0% for impossibility and 70% for

uncertainty) but also their confidence in their beliefs through the size of the probability interval. Elicitation here is based on questions such as:

> *What do you think is the percent chance that event E will occur? Please reply with a specific value or a range of values, as you see fit.*

For instance, rather than reporting a simple value of 50%, eliciting a range of probabilities allows the researcher to observe how confident the subject is in this belief. Complete ignorance is here expressed by a range as large as 0 to 100%, a relative lack of confidence by intervals such as 40 to 60%, and perfect confidence by the single figure of 50%. Given its central place in ambiguity models, the lack of confidence in beliefs should not be treated as a simple refinement of belief elicitation, but rather as a topic of interest in its own right (see Focus 5.13 for a survey of experimental designs for ambiguity).

Introspective judgements are often criticised due to the lack of incentives to reply truthfully. The beliefs can therefore be subject to a hypothetical bias and increased noise. An associated drawback is that other motives can emerge and generate a demand effect in the experiment (Trautmann and Van de Kuilen, 2014).

Proper Scoring Rules

The most popular method to measure beliefs is proper scoring rules (Brier, 1950; Good, 1952). This method, initially designed for weather forecasts, elicits subjective probabilities for uncertain events. An attractive feature of scoring rules is that the subject involved need not discuss probability. The individual rather selects a contingent payoff from a list of options, and all of the probabilistic inferences are then made by the experimenter. Under subjective expected value, a scoring rule that reveals the true subjective probability is said to be *proper*.

The quadratic scoring rule

The quadratic scoring rule is the most common procedure in incentivised belief elicitation. It was first used to infer truthful weather forecasts from experts. The quadratic scoring rule presents an event-contingent prospect to the subject. The event-contingent prospect pays $1 - (1 - y)^2$ if event E occurs and $1 - y^2$ otherwise. The subject then chooses their preferred value of y. Table 5.5 presents one quadratic scoring rule; in Table 5.5, subjects choose their preferred line.

The principle of the quadratic scoring rule in Table 5.5 is that individuals are penalised for their lack of faith in the event in a quadratic manner. If event E occurs and the respondent assigns a probability 1 to event E (selecting the last line of Table 5.5) then one unit of payment is earned. If, however, the respondent is unsure this is penalised by the probability of the complementary event E^c: $1 - y$. The penalty is $(1 - y)^2$: for a 10% lack of faith in event E, the decision-maker loses $(0.1)^2 = 0.01$ in case event E occurs. The same logic applies to event E^c: subjects are penalised for their lack of faith in a quadratic manner (y^2). Under the assumption that subjects choose y to maximise their expected value and have risk-neutral preferences, y is the subjective probability and the scoring rule is proper. The chosen y value maximises the subjective expected score.

Focus 5.7
Prediction markets

Prediction markets are a way of aggregating different opinions and heterogeneous beliefs: prices are used to predict (future) unknown events. Arrow et al. (2008, p. 877) define them as 'forums for trading contracts that yield payments based on the outcome of uncertain events'. These markets became very popular at the beginning of the 20th century, before the development of scientific polling, and served to forecast election results in the US, providing most of the time fairly accurate results (Rhode and Strumpf, 2004). Prediction markets are still popular nowadays to predict electoral or sporting outcomes. Internal corporate markets are also used by large companies to aggregate information on various internal economic indicators. The basic contract offered in a prediction market is a binary option – an all-or-nothing contract – on the occurrence of an event E. The contract costs q and pays a dollar if event E happens and 0 otherwise. Let each participant's belief $\mathcal{B}_i(E)$ be drawn from a non-degenerate continuous distribution G, and suppose that wealth w is independent of beliefs. Following Wolfers and Zitzewitz (2006), further assume that each participant chooses the optimal number of contracts y_i by maximising their subjective expected utility with a logarithmic utility function:

$$max_{y_i} EU = \mathcal{B}_i(E)\log[w + y_i(1-q)] + [1-\mathcal{B}_i(E)]\log(w - y_i q)$$

The first-order condition yields the demand for the binary option:

$$y_i = w\frac{\mathcal{B}_i(E) - q}{q(1-q)}$$

Participants choose to invest in the binary option if $\mathcal{B}_i(E) > q$ and sell it if $\mathcal{B}_i(E) < q$. The market clears when supply, $\int_q^{\infty} w\frac{q-\mathcal{B}_i(E)}{q(1-q}dG[\mathcal{B}_i(E)]$, equals demand, $\int_{-\infty}^{q} w\frac{\mathcal{B}_i(E)-q}{q(1-q}dG[\mathcal{B}_i(E)]$, and the equilibrium price is

$$q = \int_{-\infty}^{\infty} \mathcal{B}_i(E)dG(\mathcal{B}_i(E)) = \mathbb{E}[\mathcal{B}_i(E)]$$

At the equilibrium, the market price reveals the average belief among participants (if beliefs and wealth are correlated, the price is no longer the average belief but rather a weighted average, based on relative wealth). Wolfers and Zitzewitz (2006) show how different utility functions lead to different predictions for the market price q. Manski (2006) shows that when participants are risk-neutral, the market price does not necessarily reveal the average belief, but rather the $1-q$th quantile of the distribution of beliefs. In this case, prices that are close to 0 or 1 are very informative about mean beliefs. On the contrary, prices close to 1/2 are the least informative. Manski (2006) also points out that introducing risk aversion requires knowledge of participants' preferences to be able to interpret the prediction-market price. One of the great advantages of prediction markets is that they provide not only incentives for the truthful revelation of predictions but also incentives to seek out useful information to form these predictions. As uncalibrated predictions produce losses, prediction markets help participants to adjust their predictions. Prediction markets have been compared to various forms of judgement-aggregation methods, such as the Delphi method. Prediction markets were found to be a satisfactory aggregator of opinions, but are found to be difficult by participants. Also, individuals' predispositions, as, e.g., risk aversion (documented in Boulu-Reshef et al., 2016), may impede their ability to participate in prediction markets. Healy et al. (2010) show that a

double-auction prediction market performs relatively well in an environment with a simple information structure involving a binary event, even with only three players on each side of the market . For more complex environments with correlated events, prediction markets perform much less well. A common criticism of prediction markets is that they can be manipulated by placing particular orders to convince the market of the impossibility/certainty of an event. However, existing evidence finds that prediction markets are actually difficult to manipulate. Camerer (1998) attempted to manipulate betting on horse races by cancelling $500 and $1,000 bets at the last moment, but found barely any effect. Hanson et al. (2006) also show that manipulators do not distort price accuracy in the run in a controlled experiment.

Table 5.5 A quadratic scoring rule

Preferred value of y	Your payment if E is true	Your payment if E is not true
0	0.00	1.00
0.1	0.19	0.99
0.2	0.36	0.96
0.3	0.51	0.91
0.4	0.64	0.84
0.5	0.75	0.75
0.6	0.84	0.64
0.7	0.91	0.51
0.8	0.96	0.36
0.9	0.99	0.19
1	1.00	0.00

The quadratic scoring rule can be designed to elicit y to any degree of precision and to any rescaling of the payment, as shown in Figure 5.2 – adapted from the design of Offerman et al. (2009). It shows the elicitation of y to within 1% with large payments in experimental currency units. The event for which the probability is evaluated is shown on the top right of the figure: that the price of a stock at the end of the calendar year be in the area shaded purple.

The quadratic scoring rule can be extended to n events. With K possible events $(E_1, E_2, \ldots, E_K)$, a quadratic scoring rule defines a collection of scoring functions $s_i(y)$, with:

$$s_i(\mathbf{y}) = a - b(1 - y_i)^2 - b\sum_{j \neq i} y_j^2$$

with $\mathbf{y} = (y_1, ..., y_K)$ and $y_1 + \ldots + y_K = 1$. Focus 5.8 shows how the quadratic scoring rule can be adapted to measure beliefs over continuous random variables.

A more general definition of scoring rules

The quadratic scoring rule is a special case of a more general class of binary scoring rules. A binary scoring rule presents an event-contingent prospect $S(E, y) = s_1(y)_E s_2(y)$ that pays $s_1(y)$ if the event E occurs and $s_2(y)$ otherwise. The individual chooses y such that y maximises the expected value of the prospect:

$$y = Argmax\{\mathcal{B}(E).s1(y) + [1 - \mathcal{B}(E)].s2(y)\}$$

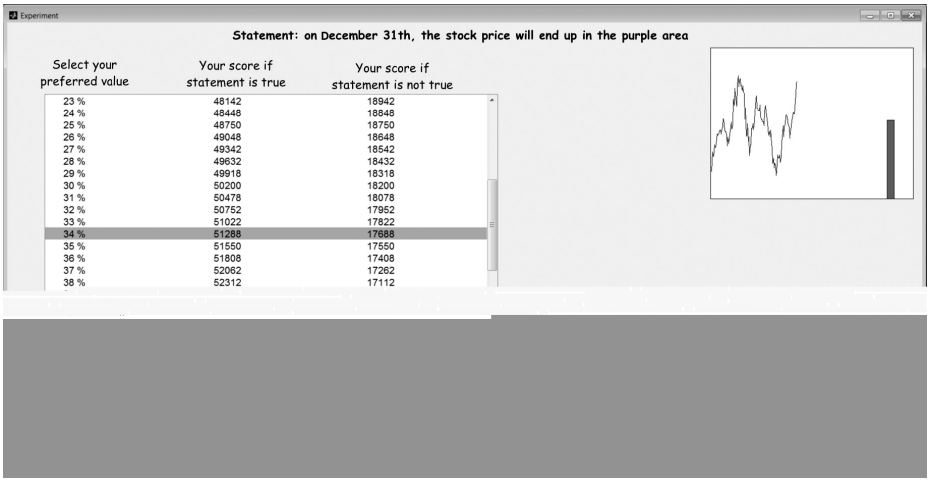


Figure 5.2 A typical display for an experimental quadratic scoring rule

Table 5.6 Examples of binary scoring rules

	$s_1(y)$	$s_2(y)$
Quadratic	$1-(1-y)^2$	$1-y^2$
Generalised binary	$a-b(1-y)^2$	$c-by^2$
Logarithmic (Toda, 1963)	$-\log(y)$	$-\log(1-y)$
Spherical (Roby, 1964)	$\frac{y}{(y^2+(1-y)^2)^{0.5}}$	$\frac{1-y}{(y^2+(1-y)^2)^{0.5}}$
Power quadratic (Selten, 1998)	ay^{a-1} $-(a-1)(y^a+(1-y)^a)$	$a(1-y)^{a-1}$ $-(a-1)(y^a+(1-y)^a)$

The first-order condition is

$$\mathcal{B}(E)=\frac{1}{1-\mathcal{A}}$$

where $\mathcal{A}$ is the ratio of the derivatives of the event-contingent payoffs:

$$\mathcal{A}=\frac{ds1(y)}{dy}\Big/\frac{ds2(y)}{dy}$$

Table 5.6 shows some of the classic binary scoring rules proposed in the literature. With this notation, the scoring rule is proper if the subjective probability maximises the payment implemented by the scoring rule (i.e. the score is maximised at $y=\mathcal{B}(E)$). If the scoring has a unique maximum it is called a strictly proper scoring rule. Savage (1971) proposes a general characterisation of strictly proper scoring rules.

Focus 5.8
Measuring beliefs over a continuous random variable

All scoring rules presented in the text are designed to measure beliefs over discrete random variables. The scoring rules can be adapted to measure beliefs over a continuous random variable with a cumulative distribution function G. The procedure starts by selecting a random divider ε that splits the continuous random variable into two groups. The probability of the left interval is $G(\varepsilon)$ and that of the right interval is $1 - G(\varepsilon)$. It is therefore possible to apply a binary scoring rule that pays $s_1(y)$ if the underlying random variable is below ε (i.e. if $G(\varepsilon)$ occurs) and $s_2(y)$ otherwise. If the subject chooses y to maximise subjective expected utility and is risk-neutral, then y is equal to $G(\varepsilon)$. Applying the procedure repeatedly for different values of ε allows the experimenter to elicit the cumulative distribution of beliefs G.

In the same vein, Schlag and van der Weele (2015) propose the following method to elicit the most likely interval for subjects' beliefs over continuous random variable with single peaked distribution: for a random variable with support $[b_l, b_h]$ and a reported interval $[b_l^i, b_h^i]$, the subject gets $(1-\frac{b_h^i-b_l^i}{b_h-b_l})^a$ if the realization of the random variable lies in the interval $[b_l^i, b_h^i]$ and nothing otherwise. This payment scheme is based on a trade-off between precision (i.e the length of the interval b_l^i, b_h^i) and the likelihood of the payoff.

Elicitation biases

A number of elicitation biases have been identified in the literature. Most of these are preference-based, referring to risk aversion, probability weighting and ambiguity attitudes. Some elicitation biases can also occur when the experimental payment and the subject's wealth are correlated.

Winkler and Murphy (1970) show that risk aversion might bias the measurement of beliefs. In addition, it might also trigger hedging in experimental games (see Focus 5.10 for details). For example, if risk-averse subjects make decisions in Table 5.5, then choosing a value y between 0.5 and the true belief $\mathcal{B}(E)$ yields higher utility. This measure is biased downward if the belief is higher than 50% and upward if the belief is lower than 50%.[7] Assume that the subject maximises the expected utility of the prospect:

$$EU = \mathcal{B}(E).u[s_1(y)] + [1 - \mathcal{B}(E)].u[s_2(y)]$$

The first-order condition is still $\mathcal{B}(E) = \frac{1}{1-\mathcal{A}_u}$, but $\mathcal{A}_u$ now also includes the marginal rate of substitution between $s_1(y)$ and $s_2(y)$:

$$\mathcal{A}_u = \frac{ds1(y)}{dy} \Big/ \frac{ds2(y)}{dy} \times \frac{u'(s_1(y))}{u'(s_2(y))} \tag{5.1}$$

Beliefs can be corrected for risk aversion in a number of ways (see, for example, the method presented in Focus 5.9). One is to measure risk aversion independently and then correct beliefs for the marginal rate of substitution between $s_1(y)$ and $s_2(y)$. Another is to use very small stakes, so that utility is approximatively linear. If, however, subjects

[7] Offerman and Palley (2016) show that loss aversion, an alternative and well-documented source of risk aversion over and above the standard curvature of the utility function, also generates a bias towards 50%.

Focus 5.9
The binarised scoring rule

Hossain and Okui (2013) propose a scoring rule that is incentive-compatible irrespective of the subject's risk preference. In a standard scoring rule, the outcome of the event E determines the size of the reward. In the binarised scoring rule, the event outcome E determines the probability of the reward. More precisely, the individual receives a fixed reward x if the payoff resulting from the draw, calculated using a loss function $L(y, E)$, is lower than a number drawn from a uniform distribution. The elicitation procedure for beliefs over an event E works as follows:

1. A loss function is defined. To binarise the quadratic scoring rule, the loss function $L(r, E)$ is $(1 - y)^2$ if E occurs, and y^2 otherwise.
2. The incentive scheme is a BDM-type mechanism. An i.i.d. draw ε is taken from a uniform distribution over [0, 1]. If ε is larger than the loss function $L(y, E)$, the participant receives the reward x. If not, the participant receives nothing.
3. The participant chooses $y = argmax\{(1 - Pr[\varepsilon \leq L(r, E)]).u(x)\}$.

Hossain and Okui (2013) provide experimental evidence showing that the binarised version of the quadratic scoring rule leads to better prediction than the standard rule, especially for risk-averse subjects. Selten et al. (1999) compare payment in lottery tickets to payments in money and find opposite results. Due to compound risk aversion, the use of lottery tickets brings about greater departures from risk neutrality.

continue to exhibit risk aversion for small stakes (Holt and Laury, 2002), small stakes will not eliminate this measurement bias. A third possibility is to pay the prize in lottery tickets (Smith, 1961, 1966; McKelvey and Page, 1990). As participants always prefer to receive a reward than no reward at all, they will maximise the probability of receiving the reward, independently of their degree of risk aversion. In other words, as expected utility implies linearity in probability, this procedure filters out risk aversion. Focus 5.9 shows how a binarised scoring rule controls for risk aversion. Contrary to Selten et al. (1999), Harrison et al. (2014) find that paying the prize in lottery tickets produces a shift towards risk neutrality. Paying lottery tickets has, however, been almost abandoned because empirically it is very nonlinear.

Another preference-based dimension biasing elicitation is probability weighting (Offerman et al., 2009). We illustrate this point by following Tversky and Wakker (1995) and Wakker (2004) in assuming that beliefs are transformed through a weighting function $\omega(\mathcal{B}(E))$, where ω is a strictly increasing probability-weighting function that maps probabilities onto [0, 1]. Under these assumptions, the subject maximises the value function:

$$V = \omega[\mathcal{B}(E)].u[s_1(y)] + [1 - \omega(\mathcal{B}(E)].u[s_2(y)]$$

Probability weighting modifies the first-order condition, which now becomes:

$$\mathcal{B}(E) = \omega^{-1}\left(\frac{1}{1 - \mathcal{A}_u}\right)$$

Focus 5.10
Risk aversion and hedging in experimental games

In experiments that include belief elicitation about the action choice of an opponent, risk aversion can lead to hedging. Here, a participant can choose an action that best responds to the action choice of their opponent, and predict the opposite action choice in the belief-elicitation task. This combination of belief elicitation and action offers insurance against low payoffs for risk-averse subjects. Hedging can then badly distort beliefs. One way of avoiding hedging is to clearly distinguish the belief and the game parts of the experiment (Costa-Gomes and Weizsäcker, 2008). Another possibility is to randomly pay either the accuracy of subjects' elicited beliefs or the payoff associated with the game's outcome (Blanco et al., 2010). Blanco et al. (2010) test for hedging in a sequential prisoners' dilemma and a coordination game. Two treatments were compared in both cases: one in which both the belief-elicitation task (a quadratic scoring rule) and the outcome from the game are paid (the HEDGING-PRONE treatment) and another in which only one task was randomly selected for real pay (HEDGING-PROOF). The following table displays the payoffs in the coordination game.

		Player 2	
		A	*B*
Player 1	*A*	(0, 0)	(16, 14)
	B	(14, 16)	(0, 0)

The payoffs are symmetric in the case of coordination on (A, A) or (B, B). Asymmetric payments move subjects away from both 50–50 beliefs and any obvious focal point. In this game the best response is to play *A* if the subjective probability about the action of the other player is over 0.533 and *B* otherwise. The mixed-strategy equilibrium is to play *A* with probability 0.533. The incentives for stating beliefs come from a linear payoff function with 10 possible degrees of belief (from 'strongly *B* rather than *A*', denoted b_5, to 'strongly *A* rather than *B*', denoted a_5). The table below reproduces the payoffs for each possible degree of belief when either *A* or *B* is the actual choice of the other player.

Belief	a_5	a_4	a_3	a_2	a_1	b_1	b_2	b_3	b_4	b_5
A	15	13	11	9	7	4	3	2	1	0
B	0	1	2	3	4	7	9	11	13	15

In HEDGING-PRONE, risk-averse subjects (with standard CRRA coefficients and reasonable 'true' beliefs close to one-half) should state beliefs with maximum certainty (a_5 or b_5) to insure themselves against risk. Moreover, choosing action *A* and belief a_5 offers a higher payoff in all cases. The combinations $(a_5,\ A)$ and $(b_5,\ B)$ are found to be significantly more frequent in HEDGING-PRONE than in HEDGING-PROOF (although this difference disappears when beliefs can only be expressed with certainty). Some players do not hedge, either due to risk neutrality or because this strategy is a best response to others' hedging. In the latter case, behaviour is consistent with higher-order hedging. A significant minority of subjects were found to hedge in HEDGING-PROOF. As such, while paying one task at random reduces hedging, it does not eliminate it entirely.

with $\mathcal{A}_u$ defined by (5.1). Both the marginal rate of substitution and probability weighting now play a role in the elicitation of subjective probability.

A third preference-based feature affecting elicitation is the attitude towards ambiguity. We have up to now implicitly assumed probabilistic sophistication (Machina and Schmeidler, 1992), so that probabilistic beliefs can be represented by a probabilistic measure. For a given probabilistic measure, this implies the same behaviour whether uncertainty is objective or subjective. One way to account for departures from expected utility without assuming probabilistic sophistication is to assume a biseparable utility model (Miyamoto, 1988; Luce, 1991; Ghirardato and Marinacci, 2001), which includes many ambiguity models as special cases.[8] Under binary RDU, the subject maximises the following value function:

$$V = W(E).u[s_1(y)] + [1 - W(E)].u[s_2(y)]$$

where W is a unique weighing function, which shares with probability measures the properties that $W(\emptyset) = 0$, $W(S) = 1$ and $W(E) \leq W(E')$ if $E \subseteq E'$ but which may be non-additive. The first-order condition is

$$W(E) = \left(\frac{1}{1 - \mathcal{A}_u}\right)$$

with $\mathcal{A}_u$ defined by Eq. (5.1). Together with the marginal rate of substitution, ambiguity attitudes now play a role through the weighting function $W(.)$. The correction of elicitation biases thus requires the measure of both risk and ambiguity attitudes.

In addition to preference-based phenomena, a possible correlation between experimental payments and the subject's wealth can distort elicitation. Take the example of an experimental game in which both beliefs and strategies are played for real. Belief elicitation and actions here are two potentially correlated sources of experimental income. A number of factors can explain this correlation. First, if the subject is risk-averse, there is a possible correlation between experimental payments and wealth. Armantier and Treich (2013) show that under expected utility, increasing the stakes in a quadratic scoring rule modifies the bias towards one-half in a way that depends on the shape of relative risk aversion. Second, both the score function and the subject's wealth may depend on the same event (a movement in a financial market, for example, or the actions of other players in a game). This correlation lowers the reported probabilities due to diminishing marginal utility. By reducing the reported probabilities, the scoring rule allows the subject to increase their final wealth in case the complementary event occurs. In this case, the scoring rule serves as a transfer scheme between the positive stake, on the one hand, and the no-stake condition on the other: an underlying condition for belief elicitation is violated (Kadane and Winkler, 1988). Last, if the decision-maker can choose some action that depends on the underlying event (e.g. an investment for which the return depends on the event outcome), there are hedging opportunities between the return from the actions and the payment from the scoring rule. In this case, hedging creates an incentive to report constant probabilities.

[8] Some examples are maxmin expected utility (Gilboa and Schmeidler, 1989), alpha-maxmin expected utility (Ghirardato et al., 2004), contraction expected utility (Gajdos et al., 2008), Choquet expected utility (Schmeidler, 1989) and prospect theory (Tversky and Kahneman, 1992).

Matching Probabilities

Decision analysis has a long tradition of using matching probabilities to elicit beliefs (Raiffa, 1968; Spetzler and Stael von Holstein, 1975). The probability p is a matching probability of an event E if the subject is indifferent between receiving an amount of money if event E occurs and receiving the same amount of money with probability p. The matching probability is defined by indifference between two prospects: an uncertain prospect paying x if E occurs and a risky prospect paying the same amount x with probability p:

$$x_E 0 \sim x_p 0 \tag{5.2}$$

When the subject is indifferent between both prospects, and under the assumption of probabilistic sophistication,[9] p is the subjective probability of the event, $\mathcal{B}(E)$. Two main procedures are used in the literature to elicit these matching probabilities. Direct procedures ask subjects to state this indifference value, while indirect procedures are based on comparative judgements or choices. As direct procedures suffer from a number of biases, we will present only the indirect procedures. Measuring matching probabilities is fairly simple with choice lists. Focus 5.11 shows how matching probabilities can be used to test complex ambiguity models.

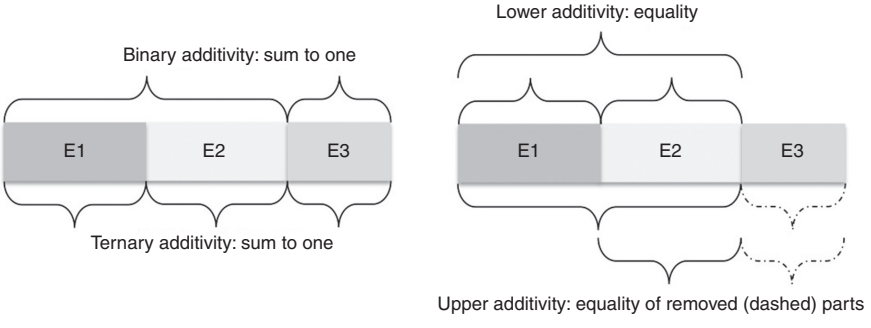
An example is the three-step procedure introduced in Baillon and Bleichrodt (2015) (see also Binmore et al., 2012). The first step uses a table similar to that in Figure 5.3. When the subject switches from the uncertain to the risky prospect between two probabilities $p\%$ and $p + 10\%$, a second table appears with probabilities $p\%$, $p + 1\%$, …, $p+10\%$ for the risky prospect. When the subject switches between probabilities $p\%$ and $p + 1\%$, then the midpoint of the switching interval $p + 0.5\%$ is taken as the matching probability of the event. A third confirmatory and pre-filled table asks the subject to confirm the choice for [against] the uncertain prospect for all probabilities below [above] their measured matching probability. Real incentives are applied in the last choice list with 101 choices. A number between 0 and 100 is selected, corresponding to a choice number. If the subject had chosen the risky prospect for that choice number, the corresponding lottery is played for real. If the subject had chosen the uncertain prospect, the observation of the event determines the payment.

Karni (2009) presents a similar procedure with real incentives. Instead of a matching probability p, the procedure elicits a number y between 0 and 1 that is compared to a uniform random draw ε over [0, 1]. If the elicited number y is greater than ε, the subject plays the lottery $x_\varepsilon 0$; if it is less than ε the subject plays the lottery $x_E 0$. Karni (2009) shows that this is equivalent to an English clock mechanism, where there is a continuous increasing bid auction between the subject and a dummy bidder. The dummy bidder stays in the auction as long as the bid is smaller than the selected number and drops out when the bid equals that number. Starting at 0, the bid increases continuously as long as the subject and the dummy bidder are both 'in the auction' and stops when one of them drops out or the bid reaches 1. Hao and Houser (2012) compare the direct declaration of

[9] Probabilistic sophistication (Machina and Schmeidler, 1995) applies if the individual uses a probability measure to determine the probability distribution over the outcomes implied by any uncertain prospect and compares prospects only by their induced probability distributions over the outcomes. This assumption goes well beyond expected-value maximisation and even expected-utility maximisation, and allows for a large class of preferences. It is not, however, compatible with the Ellsberg (1961) paradox.

Focus 5.11
Using matching probabilities to test complex ambiguity models

Baillon and Bleichrodt (2015) use matching probabilities to test the descriptive validity of ambiguity models. These models have been proposed to explain the aversion to uncertainty over probabilities (or ambiguity aversion) revealed by the Ellsberg (1961) paradox. Ambiguity aversion challenges the existence of subjective probabilities and therefore also challenges the validity of probabilistic sophistication. As matching probabilities elicit subjective probabilities only under probabilistic sophistication, it is possible to infer the descriptive validity of ambiguity models through violations of probabilistic sophistication. Different indices of the violation of probabilistic sophistication can be calculated from a threefold partition of the state space (the set of all possible states of nature). The figure below provides examples of such a partition and how the various indices are calculated. The first two indices refer to the usual additivity property: the sum of the matching probabilities for an event and its complement must be one (binary additivity), and so must the sum of the matching probabilities for the three disjoint events partitioning the state space (ternary additivity – that can be defined over single events, as shown in the figure below, but also over their complements; in the latter case, the sum of the matching probabilities over these two-event unions must sum to two). The next two indices are based on Tversky and Wakker (1995). Lower additivity implies that an event E_2 has the same impact when it is added to the null event as when it is added to a non-null event E_1. If this is the case, then the matching probability of the union $E_1 \cup E_2$ should be equal to the sum of the matching probabilities for events E_1 and E_2. Upper additivity is a little more complicated, and implies that an event E_3 has the same impact when it is subtracted from certainty (the universal event) and when it is subtracted from an event $E_2 \cup E_3$. If this is the case, then one minus the matching probability of the event E_3 subtracted from certainty (i.e. the event $E_1 \cup E_2$ in the figure below) should be equal to the difference between the matching probabilities of events $E_2 \cup E_3$ and the event E_2.



A large array of decision models under ambiguity can be tested using matching probabilities: maxmin expected utility (Gilboa and Schmeidler, 1989) and its generalisations – α-maxmin expected utility (Ghirardato et al., 2004) and the variational model (Maccheroni et al., 2006) – Choquet expected utility (Schmeidler, 1989) and its generalisation to prospect theory (Tversky and Kahneman, 1992; Wakker, 2010), and more recent models – for example the smooth ambiguity model of Klibanoff et al. (2005) and vector expected utility (Siniscalchi, 2009). Baillon and Bleichrodt (2015) find robust violations of probabilistic sophistication for both

gains and losses. Subjects overweigh unlikely events, and underweigh likely events. The latter effect is larger for losses than for gains. The violations of probabilistic sophistication form four groups of ambiguity attitudes: ambiguity seeking for unlikely gains and likely losses and ambiguity aversion for likely gains and unlikely losses. These are not compatible with ambiguity models that assume uniform ambiguity attitudes and no difference between gains and losses. Three models are compatible with the observed data: α-maxmin, Choquet expected utility and prospect theory

Which option do you prefer?

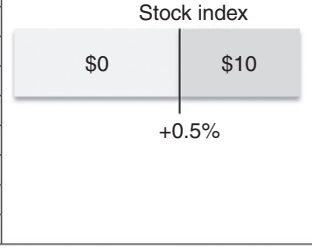
Choice number	**Option A** You win \$10 if the stock index increases by more than 0.5% (and nothing otherwise)	I choose A	I choose B	**Option B** You win \$10 with the following probability (and nothing otherwise)
1				0%
2				10%
3				20%
4				30%
5				40%
6				50%
7				60%
8				70%
9				80%
10				90%
11				100%

Figure 5.3 A typical display for eliciting matching probabilities

the number r and the bidding mechanism. The results show the continuous increasing bid auction censors naive responses and yields more accurate belief responses. One possible explanation of the difference is that the bidding procedure helped subjects think more about their answers. Another explanation is that the two-stage representation of the incentive scheme in the declarative procedure is cognitively too demanding for subjects.

Several methods for the elicitation of matching probabilities with continuous variables appear in Raiffa (1968) and Abdellaoui et al. (2016a). Consider a continuous random variable X with values over the interval $[a, b]$. The subject faces a series of binary choices between an option A that pays 1 if the underlying continuous variable is in $[a,\ m]$, and nothing otherwise (a prospect $1_{[a,\ m]}0$) and an option B that pays 1 if the underlying continuous variable is in $[m, b]$, and nothing otherwise (a prospect $1_{[m,\ b]}0$). A series of choices shift the value of m until the subject is indifferent between the two options. The indifference value m creates a partition of the interval $[a,\ b]$ into two equally likely exchangeable sub-intervals:

$$Pr[a < X < m] = Pr[m < X < b]$$

the indifference value m is the median of the distribution of beliefs. Raiffa (1968) shows how the quantiles of the distribution can be measured in the same way. Eliciting quantiles does, however, require chained answers (the previous answers are used as inputs to subsequent choices), which may lead to strategic misrepresentation and the propagation

of error. Using exchangeability, Abdellaoui et al. (2016a) show how to elicit any beta distribution of beliefs without using chained answers.[10]

Certainty Equivalents

Kadane and Winkler (1988) and Heinemann et al. (2009) use certainty equivalents to elicit beliefs.[11] Certainty equivalents are based on indifference between an uncertain prospect that pays x if E occurs and a sure prospect paying the certain amount x_c with probability 1.

$$x_E 0 \sim x_c \tag{5.3}$$

If utility is linear, then the elicited belief for event E is taken to be $\mathcal{B}(E) = \frac{x_c}{x}$. If the individual is risk-averse, then $\frac{x_c}{x}$ understates the individual's odds in favor of E. With a non-linear utility function u, the elicited belief $\mathcal{B}(E)$ is

$$\mathcal{B}(E) = \frac{u(x_c)}{u(x)}$$

Trautmann and Van de Kuilen (2014) elicit beliefs via certainty equivalents with $x = 15$. The design is based on a multiple-choice list in which the amount varies between 0 and 15 in 21 equally sized steps. Trautmann and Van de Kuilen (2014) also include an additional decision task under risk to correct the reported beliefs for utility curvature. Heinemann et al. (2009) use a similar procedure to measure strategic uncertainty in one-shot coordination games. In their design, and for each coordination game, participants were asked to choose between the safe amount x_c and an option in which the participants earn x if at least a certain percentage of the other players make the same choice, but 0 otherwise. Here x_c can be interpreted as the certainty equivalent for strategic uncertainty in the coordination game, and is used to infer beliefs over the fraction of cooperating players. Focus 5.12 compares the certainty equivalent method with others elicitation methods.

Bayesian Truth Serum

We have up to now supposed that both the experimenter and the subject can observe whether the event E occurs at some point to determine payments. This is, of course, not always the case, typically when the event refers to the subject's private information. Prelec (2004) proposes a scoring method, the 'Bayesian truth serum', for questions dealing with private information. This complements the above-mentioned methods when the truth of certain propositions cannot be checked. It is based on a scoring system that induces truthful answers from a sample of Bayesian expected value-maximising participants.

There are two questions. First, a personal question is asked:

What do you think is the percent chance that event E will occur?

[10] Neri (2015) also fits beta distributions with beliefs elicited interval by interval, which is incentivised using a scoring rule.

[11] This has also been called the promissory-notes method in the literature (De Finetti, 1974).

Focus 5.12
Comparing elicitation methods

Trautmann and Van de Kuilen (2014) compare measures of beliefs in a two-player ultimatum game using different elicitation methods in a between-subject experiment. The ultimatum-bargaining game is based on six possible proposals. For each proposal, the belief that the allocation will be rejected is elicited for the proposers and the belief that the proposal will be chosen by the proposer is elicited for the responder. The elicitation methods are the following:

- introspective judgement,
- quadratic scoring rule,
- matching probabilities,
- outcome matching.

Corrections for risk attitudes for the quadratic scoring rule and the outcome-matching procedure are also included: that in Offerman et al. (2009) for the quadratic scoring rule and that in Heinemann et al. (2009) for the outcome-matching procedure. The relative performance of the different elicitation procedures is assessed through the internal validity of the elicitation method (i.e. do reported beliefs satisfy the additivity of elicited probabilities for complementary events? Are individual choices in the game consistent with subjective beliefs?) but also with the calibration of elicited probabilities. Trautmann and Van de Kuilen (2014) find neither large nor significant differences between elicitation methods in terms of additivity of beliefs. For all methods, even after correction for risk aversion, elicited probabilities are not additive: the sum of the subjective probabilities is greater than one. Hence, there is no method, incentivised or not, that is superior to others in reducing additivity bias. Moreover, a robustness experiment shows that framing with an explicit mention of probabilities and the optimality of truthful reporting does not improve additivity for the quadratic scoring rule (see also Offerman et al., 2009). Incentivised and non-incentivised methods are different, though. Incentivised methods yield beliefs that are more consistent with proposers' choices in the ultimatum-bargaining game. While introspective judgements are associated with correct prediction in less than 20% of cases, incentivised methods predict around 30 to 40% of choices. The superiority of the quadratic scoring rule and matching probability procedure over introspective judgement is also found by Massoni et al. (2014). Last, the calibration of elicited beliefs is similar for all elicitation methods. These results are consistent with the absence of consensus in the existing literature. Sonnemans and Offerman (2001), for instance, find no difference between subjects who are paid a flat fee and those with a quadratic scoring rule – hence replicating some classical results from psychology that no difference in performance is found between subjects with a proper scoring rule incentive and subjects with no monetary incentives (Jensen and Peterson, 1973). Friedman and Massaro (1998) also find no difference between hypothetical and incentivised choices in matching probabilities, even though unpaid subjects make less effort. On the contrary, in Gächter and Renner (2010) belief accuracy in a repeated linear public game is significantly higher when beliefs are incentivised via a linear scoring rule. In addition, Rutström and Wilcox (2009) find that introspective judgements can be worse predictors of actions than beliefs elicited with proper scoring rules in asymmetric matching-pennies games. Last, in Huck and Weizsäcker (2002) a quadratic scoring rule and an outcome-matching procedure yield the same aggregate results, but the quadratic scoring rule produces more accurate predictions.

Focus 5.13
Experimental designs for ambiguity

Trautmann and Van de Kuilen (2016) review the large experimental literature on ambiguity attitudes and identify three main experimental designs that are used to study ambiguity attitudes: urns, second-order probabilities and natural sources of uncertainty. The first consists of variations on the classic Ellsberg urn experiment. In the classic two-urn Ellsberg paradox (Becker and Brownson, 1964; Fox and Tversky, 1995), subjects face two urns. A risky, or clear, urn that contains red and black balls in equal proportion and an ambiguous, or opaque, urn that also contains red and black balls, but in unknown proportion. Subjects are first asked to guess a colour on which they want to bet. This first stage is required to avoid suspicion about the composition of the unknown urn. Subjects then choose between the two urns on which they want to bet (Becker and Brownson, 1964) or provide certainty equivalents for the urns (Fox and Tversky, 1995). Trautmann et al. (2011) find more ambiguity aversion in valuation tasks, such as willingness-to-pay tasks, than in choice tasks. Choosing the risky urn (or pricing it more highly) reveals ambiguity aversion. Indifference or equal pricing reveals ambiguity neutrality. Choosing the ambiguous urn (or pricing it more highly) reveals ambiguity seeking. Using several colours or numbered balls allows the analysis of ambiguity attitudes away from the 50–50 case. For example, Becker and Brownson (1964) use urns in which the number of red and black balls is constrained by the upper and lower bounds on the number of balls of each colour (e.g. between 15 and 85 black balls in the ambiguous urn). Abdellaoui et al. (2011) use eight different colours to analyse ambiguity attitudes over different levels of likelihood: low-likelihood events correspond to bets on one (or few) colours; high-likelihood events correspond to bets on seven (or many) colours over eight. The literature usually finds ambiguity seeking for low-likelihood events and ambiguity aversion for high-likelihood events (see, e.g. Curley and Yates, 1989; Trautmann and Van de Kuilen, 2016, for references). Some experiments use numbered balls instead of colours. A second-order probability design explicitly constructs an ambiguous act. For example, Chow and Sarin (2002) use the following design to make probability unknowable: the experimenter fills a box with 11 bags, each bag containing 10 balls that are red and black. The first bag has no red balls and 10 black balls, the second has one red ball and nine black balls, and so on. Next a bag is randomly drawn and serves as the opaque urn. This compound-lottery design has also been used by Yates and Zukowski (1976) and Halevy (2007). The last design uses natural sources of uncertainty to consider ambiguity, such as temperature or weather conditions, prices in asset markets, sport and outcomes from medical scenarios. Ambiguity attitudes are revealed through choices over complementary events or source preference. Natural uncertainty is different from urns in that it does not rely on the assumption of subjects' probabilistic sophistication. Whereas probabilistic sophistication is satisfied (within the urn) for any opaque urn, this is less obvious for natural sources of uncertainty. Abdellaoui et al. (2011) present an experimental method to overcome this difficulty for natural uncertainty using a series of partitions of events into two equally likely sub-events (Baillon, 2008), allowing ambiguity attitudes for natural uncertainties to be evaluated in the same way as for 'classic' urns.

This question is similar to that used in introspective judgements. Subjects are then asked to predict the empirical distribution of answers to the first question:

What is the fraction of people giving each answer in the first question?

For each question, the probability levels are treated as different response categories. For example, to elicit beliefs to 1% precision, each question corresponds to a choice list with 101 items. Based on the answers to these two questions, a score is constructed for each respondent. Denoting x_k^i the answer to item k by individual i in question 1, and y_k^i the answer to question 2, the score for individual i is the sum of an information score and a prediction score:

$$\underbrace{\sum_k x_k^i \log \frac{\bar{x}_k}{\bar{y}_k}}_{\text{information score}} + \underbrace{\sum_k \bar{x}_k \log \frac{y_k^r}{\bar{x}_k}}_{\text{prediction score}}$$

where $\bar{x}_k$ is the frequency of answers for item k and $\bar{y}_k$ is the geometric average of the predicted frequencies for item k. The information score is the log-ratio of the average of subjects' own answers to the geometric mean of the predicted frequencies. The prediction score is a penalty proportional to the difference between the empirical distribution and the prediction.[12] Prelec (2004) gives a slightly more general formula with an extra degree of freedom to weight the prediction-error penalty. Prelec (2004) shows that truthful answers to the questions maximise the expected score for any participant who believes that others are also giving truthful answers. In other words, the score transforms survey questions into a zero-sum game in which truth-telling is a Bayesian Nash equilibrium.[13] The mechanism is based on participants using their own opinions as evidence about the distribution of opinions in the population. The 'Bayesian truth serum' rewards answers that are surprisingly common and penalises those that are surprisingly uncommon (Weaver and Prelec, 2013). One difficulty with this serum is the necessity to subsidise the zero-sum game to enforce participation. As the scoring scheme depends on all answers, and appears complex, the method can be difficult to explain in an experiment. In particular, it is not possible to explain to participants why answering truthfully is the best option. Weaver and Prelec (2013, p. 301) summarise the assumptions underlying the game as follows:

> You are most likely to maximise your earnings if you answer every item 'truthfully'. By truthfully, we mean: consider each item carefully, answer honestly, and take care to avoid mistakes . . . Remember that your Truth Score will be lower if you don't respond truthfully, so the best way to earn more money overall is to answer every item honestly.

[12] More precisely: proportional to the relative entropy.

[13] This equilibrium result relies on two critical assumptions. First, the sample of participants is sufficiently large so that one single answer does not have an influential effect on the empirical frequencies. Second, inference is impersonal: participants treat their own opinions as an impersonally informative signal about the population distribution. See Prelec (2004) for more details regarding these assumptions.

Table 5.7 The constant-sum game in Nyarko and Schotter (2002)

		Player 2	
		Green	Red
Player 1	Green	(6,2)	(3,5)
	Red	(3,5)	(5,3)

5.6.2 Belief Elicitation and Behaviour in Experiments

The Predictive Power of Elicited Beliefs

An important question in experimental games is whether elicited beliefs are related to the strategies participants play. Nyarko and Schotter (2002) answer this question in the affirmative. They construct a series of two-person constant-sum games in which beliefs are elicited (in the relevant treatments) period by period using a quadratic scoring rule. The constant-sum game is depicted in Table 5.7.

The game has a mixed-strategy equilibrium. The results of the baseline experiment, including belief elicitation and with subjects being paired with the same partner for 60 rounds, show that subjects' strategic choices in the game in Table 5.7 correspond to the best responses to their elicited beliefs 75% of the time. This result still holds if subjects are randomly matched to a new opponent each period. In addition, the results suggest that elicited beliefs are not perfectly calibrated: beliefs using the past history of play by an opponent provide greater accuracy in predicting the opponent's strategy.

A number of other studies have shown the consistency of best responses and elicited beliefs (sometimes referred to as 'stated beliefs'). Bellemare et al. (2008) find in a large representative sample of the Dutch population that proposers' belief distributions over the possible actions of responders in an ultimatum-bargaining game fit the observed choice data better than a model which assumes that proposers have rational expectations. In various finitely repeated games, Danz et al. (2012) show that actions in an asymmetric 3×3 game and beliefs are consistent. In their experiment, except for subjects with low levels of reasoning, participants with level-k beliefs consistently choose level-k actions.[14] Hyndman et al. (2013) find a similar consistency rate (63%) in a set of 12 two-person 3×3 normal-form games.[15] On the contrary, Rutström and Wilcox (2009) find that elicited beliefs are sometimes worse predictors of actions than are the beliefs estimated from an assumed belief-updating process and the observed actions of participants and their partners. Their findings suggest, however, that elicited beliefs still exhibit a predictive advantage in games with a high frequency of changes in play.[16] Costa-Gomes and Weizsäcker (2008) provide evidence from a set of 14 two-person 3×3 normal-form games, in which subjects failed to best respond to their own elicited

[14] For subjects with level-0 reasoning, Burchardi and Penczynski (2014) find that the most common elicited belief is 50%.

[15] Hyndman et al. (2013) also investigate the stability of beliefs elicited with quadratic scoring rules. They find that that elicited beliefs are fairly stable across time and for strategically equivalent games. For strategically different games, the elicited beliefs are clearly different.

[16] In one-shot games, the results are mixed.

beliefs 50.5% of the time. Moreover, the estimated underlying beliefs based on the subjects' strategies are significantly different from their elicited beliefs.

Rey-Biel (2009) finds the opposite results in a very similar experiment based on a series of ten one-shot two-person 3×3 normal-form games with a unique equilibrium in pure strategies. First, a large majority of actions are consistent with the equilibrium predictions (70.2%), whereas the figure in Costa-Gomes and Weizsäcker (2008) is only 35%. This consistency is particularly high in constant-sum games. Second, almost two-thirds of the actions (67.2%) are responses to elicited beliefs. Trautmann and Van de Kuilen (2014) find that actions in an ultimatum-bargaining game do not best respond to the elicited beliefs, whatever the procedure used to measure the latter. Higher scores are obtained with the incentivised methods (quadratic scoring rule, matching probabilities and outcome matching), but are at best equal to 35%. For introspective judgements, actions best respond to elicited beliefs in under 20% of choices. Neri (2015) considers a game with a continuous choice set – a uniform-price double-auction experiment. In this experiment, subjects' beliefs about other participants' bidding choices are elicited for different intervals of the variable of interest and incentivised with a (non-proper) quadratic scoring rule.[17] Neri (2015) finds that even though these measured beliefs explain observed bidding choices better than alternative measures of beliefs (Bayesian Nash equilibrium beliefs or empirical beliefs), only 36% of the bids are consistent with them. On the other hand, Armantier and Treich (2009) find that elicited beliefs in a first-price auction are better predictors of bids than objective probabilities.

The Effect of Belief Elicitation on Behaviour

A number of studies have investigated whether belief elicitation can alter the strategic actions of participants in experimental games. Beliefs are found to be fairly accurate in the experiment in Croson (2000), presented in Illustration 5.16. Gächter and Renner (2010) find opposite results in a repeated linear public-good game with three treatments (incentivised beliefs, non-incentivised beliefs and no belief elicitation): when beliefs are not incentivised, eliciting them does not significantly change contribution levels relative to the no-beliefs treatment. Incentivised beliefs about the average contribution level of the other group members led to higher contribution rates. Gächter and Renner (2010) explain this by incentives that may have induced subjects to coordinate to improve their gains.[18]

The effect of incentivised belief elicitation procedures in repeated linear public-good games does not generalise to other experimental environments. Croson (2000) also provides evidence on behaviour in the game in Table 5.7 both with belief elicitation using a quadratic scoring rule and with no belief elicitation. The likelihood of choosing red or green is not significantly different between treatments in the partner-matching sessions.

[17] Armantier and Treich (2009) use a similar device to elicit beliefs in a prediction contest.

[18] Rutström and Wilcox (2009) use three similar treatments – one with incentivised beliefs through a quadratic scoring rule, one with non-incentivised beliefs through introspective judgements, and one with no belief elicitation – in an asymmetric matching-pennies game and find similar treatment effects for the 'row' player, who benefits more from the asymmetry, but not for the 'column' player.

The effect of belief elicitation is to help subjects think more about the consequences of their strategy, what Croson (2000) calls 'thinking more like a game theorist'. In their experiment on first-price auctions, Armantier and Treich (2009) find that eliciting beliefs with a prediction contest rather than a quadratic scoring rule or introspective judgements did not lead to any significant treatment effect on bidding behaviour. Costa-Gomes and Weizsäcker (2008) mix the order in which subjects choose actions in various games and carry out the belief-elicitation task. Subjects' choices did not change significantly with the order of the action and the elicitation task. Similarly, there is also no order effect in Rey-Biel (2009).

Summary

Internal validity refers to the capacity of an experiment to deliver an accurate measure of a target parameter. To that end, the experiment data-generating process must be chosen in accordance with inference requirements. The data-generating process is chosen by means of the 'experimental design': the choice of all components of the microeconomic system. Proper inference requires that (i) outcome behaviour is generated by these components, and that (ii) unobserved factors are not confounding. To address the second condition, the design of experiments operationalises the identification strategies described in Chapter 3: blocking strategies amount to holding constant unobserved heterogeneity, in order to get rid of possibly confounding variations; while randomisation breaks any correlation between unobserved components of the outcome variable and the variations of interest. Several components of the experimental design aim to strengthen the link between outcome behaviour and the chosen dimensions of the experiment. First, monetary incentives provide control over preferences, hence aligning individual motives with the chosen incentives. Second, the choice of parameters decide on the setting of the environment: control parameters remain constant in all instances of the experiment, while treatment parameters aim to deliver identifying variations in outcome behaviour. Third, the actual data-generating process is mediated by how subjects perceive the experiment, described thanks to experimental instructions, and their opponents when interactions are repeated. Subjects' beliefs are key to the internal validity of experiments, and their measurement is the matter of this chapter's case study. The next chapter turns to the actual implementation of the experiment, and describes all the practical steps from building a laboratory to the time of the experimental sessions.

6 Conducting an Experiment

This chapter describes step by step how to conduct an experiment. Most of the chapter will focus on the standard practice of computerised experiments. An alternative implementation – which is also easier to implement and does not require as much explanations – is to use pen and paper: subjects make decisions on sheets of paper that are distributed to them. In this case, the experimenter essentially does manually all the operations that are otherwise performed by the computer.

To implement a computerised experiment, the first requirement is an experimental laboratory. Setting up a lab is an important decision, and a costly investment for a research group or institution. Section 6.1 describes the main components to be taken into account when creating a lab. Once the lab is set up, it is time to consider the structure of the experiment you would like to run. The main issues relating to this structure are described in Section 6.2. Ideally these issues should be settled at least two months before the beginning of the experiment. Based on this structure, Sections 6.3 and 6.4 present the final settings and fine tuning required in the month preceding your experiment. Section 6.5 provides reminders for D-Day. Last, Section 6.6 reviews the tools used to measure time preferences – a topic regarding which implementation is generally quite challenging.

6.1 A Long, Long Time Beforehand: Setting Up an Experimental Laboratory

Setting up an experimental lab is an important investment, and involves a number of prerequisites. First, you will need to find a suitable place for the lab. Ideally, the lab will comprise a number of rooms: a waiting room, the experimental room, a room for the experimenters, a room for the hardware, and a storage room. Figure 6.1 depicts a typical set-up. In each room, you will have to pay attention to a series of characteristics. First of all, space is key, and is often a considerable constraint on experimental labs. Other characteristics such as lighting, temperature, ventilation and equipment are also of great importance and should not be neglected. Last, setting up an experimental lab also requires planning the financial procedures to pay subjects well in advance.

6.1.1 The Waiting Room

When subjects come to the lab to participate in an experiment, they often have to wait a little while until the experiment starts. A well-identified waiting area or, even better, a waiting room, can be of great help. For the experimenter, a well-identified waiting area

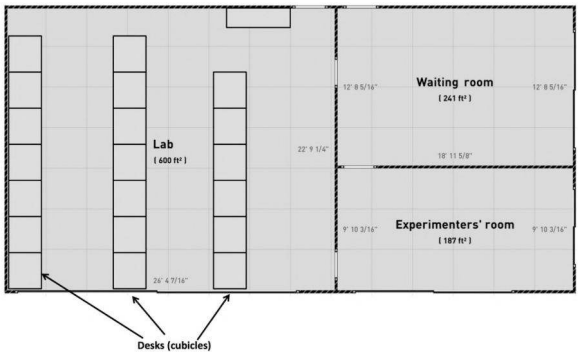


Figure 6.1 Typical implementation of an experimental lab

avoids losing time by keeping the participants together before they enter the experimental room. It is also useful for the formalities of checking in and oral announcements. A well-identified waiting area also prevents participants from getting lost and spending time looking for the lab: the waiting area is a physical focal point for the lab. Last, when deciding on the lab settings, it should be ensured that the waiting area should not generate, or suffer from, externalities. Negative externalities can arise if the waiting area is situated somewhere where silence is important: near classrooms, offices, quiet study areas, the library, etc. twenty people who are waiting for an experiment, even if only for three minutes, can generate a lot of noise and disturb the environment. Negative externalities can also pass from the environment to the subjects. This will be the case if the waiting place is near food and drink areas, cafeterias, group study places or other busy and noisy settings. If you have sufficient space and decide to set up a small-scale lab with a limited number of seats, a dedicated waiting room can be a very good option to avoid these negative externalities. For large groups, a well-identified open space is best. Whatever the option, do not forget to indicate clearly in the invitations the location of the waiting area and put up signs so that subjects can easily find the laboratory.

6.1.2 The Experimental Lab

The most important room is, of course, the experimental lab itself. Ideally the laboratory will be dedicated to research, so that the experimenter can run an experiment as soon as the recruitment process is complete. If the lab is also used for teaching, it might only be available at times that are less convenient for subjects, giving rise to recruitment

difficulties. In addition, a dedicated research lab gives the experimenter the freedom to design the space, and will save time by not having to reconfigure the lab every time it is needed.

When setting up the experimental room, the objective is to build and design an environment that is useful, safe, comfortable and efficient. The first concern will be space: this should be adapted to the maximum number of participants you expect in experiments. A spacious room is always best: a crowded lab might influence the quality of the experiments through congestion externalities. As a rule of thumb, you can consider that 25 square feet (2.5 square meters) per seat is a good compromise between space and efficiency. Using this rule of thumb, a 20-seat lab will require a 500-square-foot (50-square-meter) room. Once the necessary space is determined, it is time to consider the room's other architectural characteristics. Air conditioning and heating may be required to regulate the temperature during the experiment. Ventilation may also be needed to ensure a good atmosphere in the lab when all participants are present and are working on the computers. A good uniform light source is also important for participants to perform the experimental tasks in good conditions. Many experiments are run on computers, tablets or other electronic devices: the lighting, electricity and ventilation should be adapted to this kind of equipment. Last but not least, ensure that the room satisfies all safety requirements.

Once the experimental room has been set up, you will have to install work surfaces (tables, desks and chairs) for the participants and the experimenter. The desks should be spacious enough that the subjects are comfortable during the experiment. Comfort comes at a price: the larger the desks the fewer of them will fit into the room. Here also, you will have to think like an architect in order to pick out the optimal solution. Having desks on wheels helps if they need to be moved. Partitions or cubicles are used in most labs to separate work surfaces. The types of partition vary according to the objective, and there are a number of options. A first possibility is to install partitions to avoid visual contact between subjects during the experiment. Another possibility is to have partitions that only block the screens where the entries are made. Being able to move the partitions is important as well. Moving or adapting partitions provides greater flexibility in experiments. In this case, the partitions should ideally be light and easy to store. The number of computers in the room depends on the experiments to be run. Some experiments require only few computers while others need many. Of course, the greater the number of computers or displays, the more flexibility you have. In general, there will be one computer per seat. Computers can be standard desktop PCs, portable computers or tablets. The great advantage of desktop computers is that no specific handling or connections are necessary once they are installed. Mobile computers or tablets offer more flexibility, and the ability to have a mobile lab, but the cost of handling them can be much higher.

For flexibility, you might consider installing a number of whiteboards and at least one projector in the lab. If you wish subjects to be able to look at a whiteboard or at the experimenter during the experiment, the partitions might require you to use a different desk layout (e.g. semi-circular) than if you do not. You might also wish to place desks

in such a way that you can walk behind the subjects in a casual manner. This should be taken into account when estimating the necessary space.

6.1.3 The Experimenter's Room

It is useful to have a place where the experimenter can monitor the experiment without influencing or disturbing the subjects. If space is an issue, the experimenter's room can be replaced by a separate area in the experimental lab. The experimenter's room should be large enough to fit in a desk with a computer, a printer and/or a photocopier for the last-minute printing of instructions, receipts and cheques. If there is enough space, it can also be the place where the servers are stored. Some recommend the use of up to three servers – one to run experiments, a Web server for recruiting and the webpage, and one as a backup. However, the number of servers is often restricted to one or two for cost reasons in experimental labs.

The experimenter's room can also serve as a separate cashier's office. The ideal configuration is for the experimenter's room to communicate with both the experimental lab and the exit. As such, subjects can either exit the experimental lab directly or via the experimenter's room. A window between the experimenter's room and the lab may be a plus in order to monitor the experiment without interfering with the subjects.

Figure 6.2 An experimental lab: what it looks like

Last, it is useful to provide storage space in the experimenter's room. Many different kinds of things need to be stored in the course of running experiments: computer pieces, papers, pencils, payment materials, etc.

6.1.4 Financial Procedures

The financial procedures for subject payment in general follow two different sets of rules. First, both payment and recruitment procedures have to comply with general ethical guidelines and principles and the approval from an institutional ethics committee (IEC) and institutional review board (IRB).[1] For example, boards may require a minimum level of compensation for experiments on an hourly basis. Second, your payment procedures will depend on your institution's financial rules. Financial procedures can be very strict in terms of the maximum amount to be paid to an individual, the maximum amount per session or per period (week, month), the form of and delays in payment, etc. Payment can be made in many different ways: cash, cheques, and bank transfers or online payments, to name a few. All institutions are different and you should expect some bargaining with your own institution's finance or accounting department when setting up the lab. It is often hard to find a way for teaching and research institutions to pay students (without the cumbersome paperwork associated with formal work contracts). Cheques and cash remain the most widely used forms of payment. Some universities only allow payment by cheque, while others allow cash or electronic payment. Some universities allow the researcher to request a cash advance to pay subjects directly, provided that there is a sufficient delay between the demand and delivery, while others only allow the reimbursement of cash payments after the experiment. Payment procedures can be hard to fix or change if you have particular needs, such as delayed payments resulting from choice over time, payments to people outside the university, or payments above the maximum allowable amount.

6.2 Two Months Before: The Basics

Two months before the experiment, you can set out its basic structure. Four elements are crucial here: the instructions, the script, the subject pool and money.

6.2.1 The Instructions

As explained in Chapter 5, Section 5.4.1, the written instructions are essential to the experiment's success. To be convinced of this, just think about the many times you have tried to assemble a new piece of furniture and failed to do so properly as the instructions were not good enough. The same can happen in economic experiments. The main difficulty in writing instructions is the natural tendency to assume that the subject is an

[1] The set-up of such boards varies a great deal across locations – it can typically be part of the university as a whole, or be created by academic staff within a department. In all cases, a common practice for laboratory experiments in economics (which are typically innocuous in terms of ethics) is to define a set of requests associated with standard experiments, and grant approval automatically to all proposals matching this standardised list.

expert in economics and knows all about the experiment you are running. While most instructions have the same structure, they are fundamentally experiment-specific. As a reference, the list below summarises the main steps of a typical sequence, which is applied to the dictator game in Illustration 6.1. In Section 6.6, Illustration 6.4 shows a set of written instructions used in an individual decision-making experimental task for choice over time.

- First, welcome the participants.
- Second, set out the general rules for the experiment. These include the logic of the incentives (i.e. explaining the link between individual payment and performance) and the general principles of the experiment: no deception, no judgement of answers, anonymity and so on.
- Third, present the general structure of the experiment. This part describes the general environment (the general conditions, the number of subjects and the duration of the experiment), the nature of the interactions subjects will have with others (if any), the timing of the experiment, the general conditions linked to participation, etc.
- Fourth, describe the tasks that the subjects will have to perform during the experiment.
- Fifth, give further experimental details: the number of rounds, the roles individuals play in the experiment, etc.
- Sixth, ask subjects to fill in the pre-experiment questionnaire to check that they have correctly understood the task.
- Last, describe the way in which subjects can leave the experimental room.

6.2.2 The Scripts

Once the experimental design corresponding to your research project has been set out in the instructions, it is time to implement it via a computerised script. One simple way to start writing down a script is to write a first draft of the future instructions. By doing so, you will force yourself to have a formalised view of the experimental design and set out the logic of interaction and tasks in your experiment. This phase actually corresponds to the construction of an algorithm. Having an algorithm representing the experiment as a whole is extremely useful in order to understand the sequence of tasks that subjects will face during the experiment. Writing a script is the same as setting out the set of rules to be followed during the experiment.

The general principles of elementary computing can be used to give some minimal structure to your experimental algorithm. There are typically four basic requirements. First, the set of rules must be finite: subjects must face a sequence of tasks that can be completed in a finite time. This sounds obvious, but for instance the infinite-horizon environments that are assumed in many economic models contradict this rule. In search experiments, individuals who do not find an exchange opportunity are supposed to carry on searching forever. Game theory suggests using random termination of the task to mimic the infinite-time horizon. Alternatively, fixed but unknown (to the subjects) horizons can also be used (see Section 4.4.2 for a detailed discussion). Second, the rules must be followed in a particular order. Here the experimenter must have a clear idea of

Illustration 6.1
Experimental instructions for a simple dictator game

This illustration provides sample instructions for a simple dictator game, inspired by Hoffman et al. (1994).

1. *Welcome the participants*: Thank you for participating in our experiment on decision-making.
2. *General description of the experiment*: During this experiment, you will have to make decisions involving various amounts of money. The amount you will earn will depend on your own decisions as well as the decisions of other participants. All your responses will be converted into anonymous data after the experiment. During the experiment you must answer a series of questions. There are no right or wrong answers to these questions.
3. *General structure of the experiment*: *N* people will participate in this experimental session. For reasons of anonymity, you will not know the other participants' identities.
4. *Description of the tasks subjects will have to perform during the experiment*: In this experiment you will be paired with another person in the room. One of you will be the seller and the other the buyer. The seller chooses the selling price between 0 and \$10 and the buyer has to buy at that price.
5. *Further details about the experiment*: The following table shows the possible values of profits for the buyer and seller:

Chosen price	\$0	\$1	\$2	\$3	\$4	\$5	\$6	\$7	\$8	\$9	\$10
Seller's profit	\$0	\$1	\$2	\$3	\$4	\$5	\$6	\$7	\$8	\$9	\$10
Buyer's profit	\$10	\$9	\$8	\$7	\$6	\$5	\$4	\$3	\$2	\$1	\$0

6. *Pre-experiment questionnaire to check that the subjects have correctly understood the tasks*: If the seller chooses \$8, the seller will be paid \$8 and the buyer will be paid \$2. Do you agree (Y/N)? If the seller chooses \$2, the seller will be paid \$2 and the buyer will be paid \$8. Do you agree (Y/N)?
7. *Description of the way subjects can leave the experimental room*: Once the sellers have made their choices, the profits will be paid to participants and the experimental session will end.

not only the sequence of the tasks but also the hierarchy to be specified between them. Third, one, and only one, rule may be obeyed at a time. This requires decomposing the experiment into a series of simple basic tasks and elements. Last, the rules must cover any awkward situation that can arise during the experiment. This is crucial for the internal validity of the experiment. Figure 6.3 shows an example of a simple experimental algorithm based on the dictator game.

Most experiments are computerised, and scripts have to be written. If the experiment is not computerised, a framework for the paper-and-pencil or oral questionnaire has to be designed. Computerised experiments have a number of attractive features compared

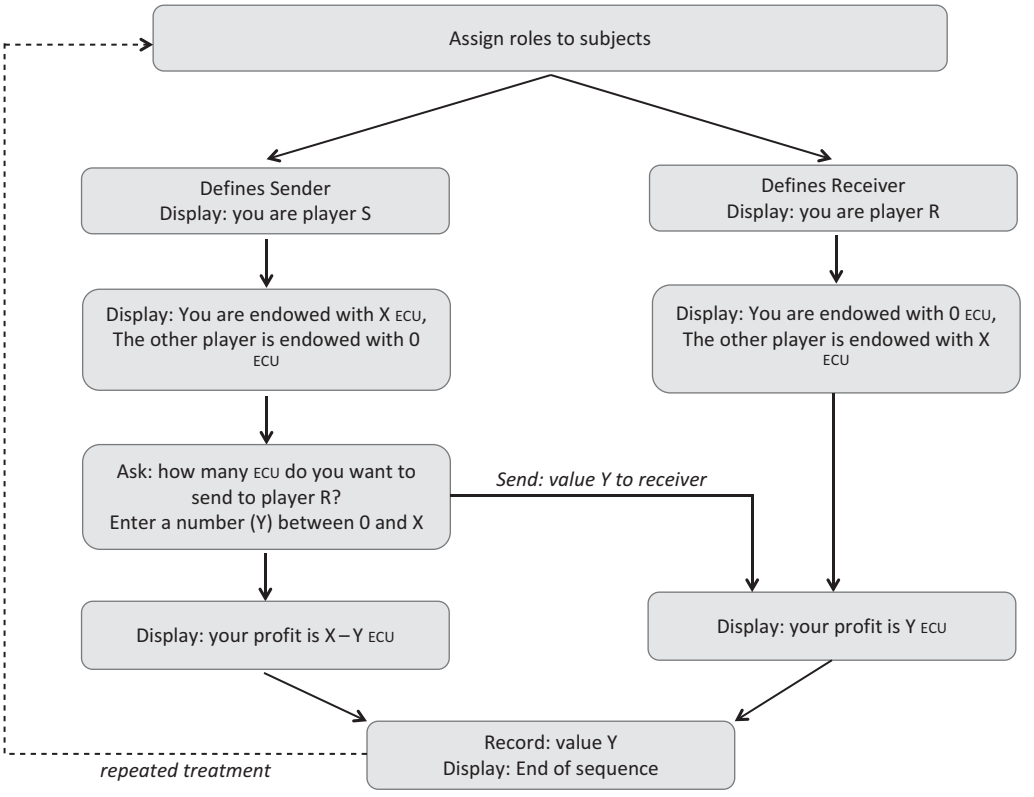


Figure 6.3 A basic experimental algorithm based on the dictator game

to paper-and-pencil experiments. First, they simplify the data collection. Second, they are fast: once the experimental set-up exists, they are quick to run and reduce the subjects' manipulation of experimental material. Third, they reduce the cost of replication. Last, they permit simple variations on a theme for experimental designs. On the other hand, running computerised experiments often requires experimental labs, with all the equipment described in the previous section. However, broad access to the Internet on different types of hardware reduces the burden of a fully equipped experimental lab. A variety of languages exists to write scripts for computerised experiments. The most common are Python (for standard apps) and PHP/MySQL combinations for Web interfaces. Both require some programming skills in order to construct a computerised experiment from scratch.

Choosing software to run the experiment is a long-term investment. As an economist you know that you should follow the rules of optimum investment choice here. To make the right choice you have to answer a series of questions. The first concerns your time horizon. If you have a short-term objective, or plan only a few experiments, you cannot afford to invest time in high-level programming, unless this can be obtained at low cost. Using a pre-programmed experiment is clearly a good option here. If you have

long-term objectives and a number of experiments planned, you might want to move to more sophisticated experimental-economics or general-purpose software. The second question concerns the type of experiment to be run. If you want to run experiments based on existing well-known designs (bargaining and sharing, auctions and simple games) almost all the software we mention below has pre-existing routines for basic experiments. Here, the simpler the better. However, if you need to run sophisticated or specialised experiments, you will need to invest in the programming of scripts. One important aspect of the choice of software is its ability to fit in with your ecosystem. The more users of a given software you have around you, the greater the incentive for you to use it as well. Your ecosystem depends on your research interests, your research community, your institution, and even your friends. The more topics you can share, the higher your productivity will be. Typically, it is inefficient to choose software requiring knowledge of Java if you do not have these skills yourself or around you.

Writing a script can take from three days to two weeks of programming. Here, two weeks of programming does not necessarily mean two weeks in calendar time, as the programming may well be spread out over a longer time period. It is not uncommon to encounter difficulties when programming: bugs, incompatibilities and logic problems. Any of these can take a significant amount of time to solve, either because it is difficult to figure out what has gone wrong or because you need help to solve the problem. This has to be taken into account when scheduling the experiment.

If you have no or few programming skills, the VeconLab pre-programmed experiments proposed by Charles Holt at the University of Virginia can be of great help. The website offers over 60 on-line programmes that can be used for teaching experimental economics and research. These include a variety of options and cover most fields in experimental economics: auctions, bargaining, decisions, finance, games, markets, public good games and various surveys. The programs are based on a PHP/MySQL combination and can be run in any Web browser. Other alternatives are Econport, proposed by the University of Arizona and Georgia State University, and Comlab. Econport is mainly teaching-oriented. It allows different types of pre-existing experiments to be combined through a Web-based interface. The main elements include basic games, consumer-utility maximisation, markets and auctions, and supply and demand graphing. Comlab is an entirely game-oriented stand-alone application with three modules: one for game design, one to run the experiment (executing the game) and a third for analysing results. If you are instead ready to invest in programming, there are a large variety of options. The leading software in experimental economics is Z-tree (Fischbacher, 2007). Z-tree offers an integrated framework, including post-experimental questionnaires and payment. One advantage of Z-tree is that it requires no prior knowledge or experience in programming. This does not mean that you will not have to do any programming, but rather that you do not have to be trained as a programmer to write a script. The logic of Z-tree makes the programming natural, as it follows what we called the experimental algorithm in Figure 6.3 rather than standard programming logic. Z-tree is a client–server application. This means that you, the experimenter, will interact with the server application (Z-tree) while the participants interact with the client applications (called Z-leaf). Note that Z-tree is only available on PC.

For more specialised topics, Caltech has developed a number of specific open-source Java-based software packages for experimental markets and auctions. Among these, the jMarkets software offers high-level options for a large variety of market configurations and jAuctions various ways of programming experimental auctions. Caltech and UCLA have also developed a more general open-source Java-based framework called Multistage, which requires Java programming experience.

Alternatives based on Python frameworks also exist. These alternatives require an understanding of Python and HTML programming, which is much less complicated than Java programming. We present two examples of these alternatives. The first example is oTree (Chen et al., 2016). In oTree the experiment code runs on a server computer (a local laptop, or a cloud server, for example) and displays the experimental scenario on the subject's devices with a Web browser. The subject's device can be a standard computer, a laptop, a tablet or a mobile phone. No specific installation is needed on the subject's device. This feature makes Python-based frameworks very flexible for use in the lab or the field, in classrooms or online, on platforms such as Amazon's Mechanical Turk or Qualtrics. The second example is Willow, developed at the Center for the Study of Neuroeconomics at George Mason University. The Willow user interface for subjects is also displayed inside a browser connected to a Web server that runs on the experimenter's computer. Both oTree and Willow offer a compromise between all-in-one integrated software and purpose-written software with standard languages.

Last but not least, software and hardware compatibility needs to be checked. A first check is that the script is internally consistent. Errors or mistakes in the design or the source code may produce unexpected behaviour or results, crashes, or the freezing of the software. The detection of such bugs and inconsistencies in the script can take a long time, which should not be underestimated when planning the experiment. Second, you need to check that your program and system are safe. Your script should not introduce vulnerabilities into your system and open gates to malicious software such as spyware, trojans or viruses. Once your script is consistent and safe, check that it runs satisfactorily on a different computer. Three kinds of incompatibility can occur here. First, you might find missing links. This happens if your script calls external files (data files, images, videos or text documents) that are not present on the computer. Second, you may experience interface incompatibility. This happens if the computer has a new version of the software that you are using to run the scripts, or if the version of the operating system is different. Last, you might experience incompatibility issues. This can happen if computers have different hardware, but also if the presence of some program on the new computer inhibits the performance of your script. While the former is entirely predictable, the latter is not. This calls for a careful check of all the environments in which you plan to deploy your script a long time before the experiment. Compatibility problems extend to computer networks. In this case, the environment to be checked is not a single computer but rather the network. Incompatibilities turn up here for the same reasons: missing links, and interface or performance compatibility. Communication protocols and data exchanges between the server and the terminal thus also need to be checked for software compatibility.

6.2.3 The Subject Pool

Two months before the experiments is not only the time to write the script(s) but also when you should ensure that you will have access to a subject pool, i.e. a population of individuals from which the participants in the experiments can be recruited. Recruiting, organising and maintaining a subject pool is a difficult task. There are a number of constraints and potential difficulties, and you will have to pay attention to a lot of small and medium-sized details to create and maintain a satisfactory subject pool.

First, before creating a subject pool, you will have to ensure that your research and experiment are not restricted by the ethical or legal procedures at your university or institution and in your country. Experiments in economics involve human subjects, i.e. individuals about whom the researcher wishes to obtain data and private information. The experimental protocol is therefore a data and information technology that may be subject to private or public regulations. In economics, the use of human subjects is simple because it is mainly based on information: most of the time, the data you require are private information obtained by communication between the subject and the researcher. For a flavour of the kind of regulation or guidance that exists for the use of human subjects, the US Department of Health and Human Services provides general and very informative guidance on the engagement of institutions in human-subject research. When recruiting the subject pool, you will have to provide the prospective participants in your experiment with some information on the general framework of economic experiments and the use of the information you will obtain from them – as detailed in Illustrations 6.2 and 6.3. This includes the general purpose of the research, the possibility of contacting researchers for information about enrolment, the institution organising the research, and the general rules of the experimental lab. Prospective subjects must also be aware that if they accept being part of the subject pool, they allow researchers to contact them for forthcoming experiments.

For example, if your subject pool is entirely based on university students, you will have to check their timetable to avoid university holidays, the exam period and any other period when students are unavailable. This results in a large number of small constraints to combine and monitor at the same time. Fortunately, some software has been developed in recent years to assist the organisation of the experimental agenda.

One of the most widely used pieces of software in experimental economics is ORSEE, a Web-based Online Recruitment System developed by Greiner (2015). Other participant-management software such as SONA can also be used to manage subject recruitment. All are either designed for the easy organisation of economic experiments or can be adapted for that purpose. Participant-management software simplifies the organisation of experiments by taking care of a number of time-consuming tasks: information and statistics about the subject pool, the standardisation of the procedures related to organisation, and the reduction of experimenter/subject interactions. Participant-management software has three main components: a public area accessible to any visitor, a private area for registered participants, and the administrator area. In ORSEE the public area allows the display of general rules (legal

Illustration 6.2
Information provided to prospective participants in economic experiments

General rules:

- Participation in experiments requires no prior knowledge in economics.
- Registration is open to everyone (students, employees, people looking for a job, the retired, etc.) without restrictions.
- Individuals can only register once.
- You can unsubscribe from the mailing list by sending an e-mail to xxx@y.zz

Invitation to participate in an experiment:

- When an experiment is scheduled, invitations are sent by e-mail.
- Only registered individuals will receive invitations.
- All registered individuals are not invited to each experiment.
- Receiving an invitation by e-mail does not guarantee participation in the experiment. The invitation e-mail contains a link to a page for enrolment. Individuals need to be enrolled to participate in the experiment.
- In general, enrolment exceeds the number of participants required for the experiment. Those who arrive on time and can participate will be compensated.
- After enrolment, a confirmation e-mail will be sent. This e-mail notes the name, date and time of the experiment.
- A reminder e-mail will be sent a few days before the experiment.
- You can cancel your enrolment by sending an e-mail to xxx@y.zz. Please specify the name, date and time of the experiment.
- If you do not show up to an experiment, you will receive an e-mail. If you do not show up three times, you will be excluded from any future invitations.

requirements, privacy policy, specific rules of the institution and payment conditions). It also includes an FAQ page and a calendar containing an overview of the experimental sessions. The calendar allows potential subjects to see the name of the experiment and some basic details: date, time, duration, location and number of free places available.

To construct a subject pool, invitations are sent to potential subjects with a link to the server, which will create and organise the subject pool. Importantly, using participant-management software confronts subjects with a standardised environment (which is important for internal validity) and a convenient way of registering for the mailing list (which is important to simplify registration). ORSEE also allows subjects to manage their registrations in experimental sessions. To complete registration, subjects only have to enter their first and last names and an e-mail address. Additional, optional, personal data can be entered in the registration form (such as gender and field of study). Once subjects have completed the registration process, ORSEE updates the participant table. Subjects have access to a similar environment in case they want to change their data or unsubscribe from the participant list. Once a subscriber is registered, ORSEE sends

Illustration 6.3
Information provided to prospective participants in economic experiments (continued)

Participation in an experiment:

- Rules to follow before the start of the experiment
  - Participants in the experiment are asked to show up in front of Room xxx a few minutes before the experiment starts.
  - An experimenter will take a roll call.
  - The experiment begins at the time specified in the invitation e-mail.
- Rules to follow when entering the experimental room
  - Switch off your mobile phone, so as not to interfere with the experiment.
  - Sit in front of one of the computers.
- Rules to follow during the experiment
  - No communication (unless otherwise instructed) with other participants in the experiment.
- Rules to follow at the end of the experiment
  - Payment of earnings in the experiment will be made individually.
  - Payment will be made by cheque, according to the rules set out in the experimental instructions.
  - Every experiment has its own compensation scheme. No information on the remuneration of the experiment is provided before the start of the experiment.
  - Most experiments include the payment from the experiment plus a participation fee of 3 euros.

Data privacy: The private information in participants' registration is only used to organise the experiments. In accordance with French Law on the protection of individuals, you may choose to modify and/or delete your personal data.
Consent: I have read the registration conditions and the privacy and data reporting and agree with the content. Yes/No

them an invitation to participate in an experiment by e-mail. This e-mail contains a link to the individual's experiment registration page. Each potential subject's page shows the experiments to which the subject has been invited. For each of these, the subject can choose to register by a simple click on a button. The page also shows the forthcoming experiments for which the subject has already registered, and those in which they have already participated.

From the experimenter's side, the software offers a number of administrative tools to manage the subject pool and lab organisation. An important function is the maintenance of the subject's pool. Different tasks are performed here: adding a new participant to the database; searching for duplicates in the database; deleting, unsubscribing and excluding participants; and modifying or deleting subjects' personal data. A second main function is the experimental calendar, which is an essential time-saving device in lab management. An experimental calendar contains all the information on lab or

subject-pool availability, bookings (date, time and location), the details of each experiment, the number of subjects who have already registered and so on. In practice, an experimental calendar can serve as the main interface for managing the lab. This is, for example, the case with ORSEE. With the experimental calendar, you can easily check subject availability and book the lab for forthcoming experiments. Moreover, the statistics, for example on the number of individual no-shows, provide information on the reliability of the subject population.

6.2.4 Money

Two months before the experiment is also the time to establish a provisional budget for your experiment, by listing the different expenses you will face and the funds available (research funds, grants, etc.). The expenses include experimental payments (show-up fees and experimental winnings), staff costs (wages and salaries to be paid to research assistants, taxes and insurance), operational costs (printing, specific material). Preparing the budget includes listing the institutional procedures and bureaucratic requirements you will face to cover the experiment's costs and pay subjects. All this needs to be established long before the experiment. In particular, this will allow you to ensure that your experiment complies with all institutional financial rules.

The budget will require you to predict total payments to subjects. Predicting experimental gains is not easy. You first need to choose an exchange rate between the experimental currency units and money. Dominance (see Chapter 5, Section 5.2.1) requires the average gain to exceed the subjects' opportunity cost of time – which varies according to the composition of the subject pool. A rule of thumb is to double the hourly wage rate (for students €10–12 *per* hour) or to refer to an hourly cost of time (e.g. the value of volunteer time is between $20 and $30 per hour in most of the US). Second, you have to be sure that all possible experimental gains are covered. For example, if you use a random-task incentive procedure, the sample distribution of real payments can be very skewed to the right (or to the left) as compared to the expected distribution.

6.3 One Month Before: The Final Settings

Now that you have established the basic structure of your experiment (scripts, subject pool and money), it is time to determine the final settings. Three elements are crucial here: planning, the pilot and the feedback from the pilot.

6.3.1 Planning

Here are the main steps to follow in order to plan and organise experimental sessions.

1. Choose the number of subjects needed for the experiment. The number of subjects depends on several factors: the number of between-subject conditions and treatments you wish to run, the statistical power of the difference between conditions, the

availability of the underlying population you are sampling from, the overall monetary and time costs of the experiment, etc. Your final choice will be a compromise between these factors.

2. Determine the number of sessions you need to run to obtain the number of participants you require. The number of sessions mainly depends on the number of subjects and the constraints of the experimental lab (e.g. the number of seats).
3. Determine the maximum duration of the experiment. Most software for the management of subject-participant pools allows the easy scheduling of the time and date of the experimental sessions. The pilot will be of great help in evaluating the minimum, mean and maximum duration of each experimental session.
4. Plan the pilot sessions, with and/or without monetary payoffs. Pilot sessions without monetary payoffs are easy to carry out, and can be used to fine-tune the design at earlier stages of the project. Pilot sessions with monetary payoffs, however, deliver a more robust test, and allow you in particular to check whether payment procedures work well and comply with all financial requirements.

6.3.2 Running the Pilot

On the practical side the pilot is your trial session. The pilot has to be as similar as possible to the real session (see Section 6.5 and Figure 6.4 for details).

The pilot will answer a number of questions. First, the pilot reveals whether the instructions are understandable and have indeed been understood by the participants. If something is badly wrong in the instructions, for instance if subjects fail to understand the task or if the tasks are too complex for the average subject, this will become apparent in the pilot session. This will prevent you from spoiling your experiment via bad phrasing, for example. Second, pilots of computerised experiments uncover unexpected bugs or errors in the software. This may concern everything from the displays provided to subjects to data encoding. Third, a pilot provides useful information on experiment duration, the timeline for sending out invitations and reminders to participants, the sample variation in subjects' hourly payments, and so on. Last, the pilot session will produce preliminary, although unreliable, results regarding behaviour. If the pilot includes payments, it will also allow you to test your payment procedures.

6.3.3 Feedback from the Pilot

The main purpose of the feedback from the pilot experiments is to improve the experimental design and to fix any remaining flaws in the implementation. Since the pilot is a small-scale version of the experiment, which may well be changed following the pilot, no inference on behaviour can be drawn from the observations here. Rather, the feedback from the pilot provides useful information on how to fix problems and prevent any unwelcome consequences from previously unforeseen events. The first piece of feedback is your ability to run the experiment. The pilot will confirm if you are ready in terms of script and software compatibility, planning and logistics. The second relates to the subjects. The subjects' reaction to the instructions will confirm if they fit well with the design of the experiment and whether any changes are necessary. The pilot will also confirm whether your recruitment procedure fits the experiment or needs to be changed

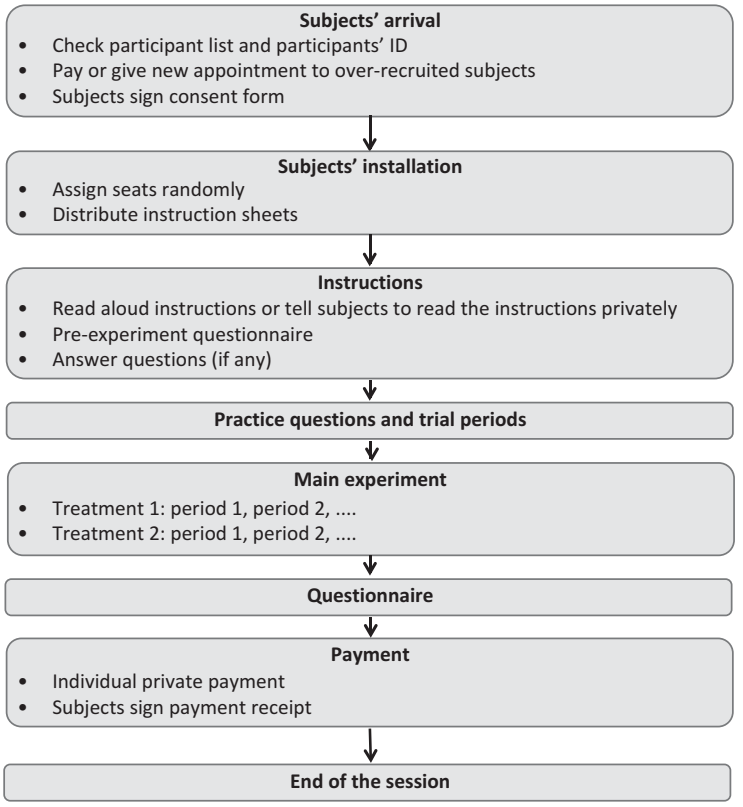


Figure 6.4 A typical experimental session

(over-recruitment, composition of the pool of subjects, or participant diary constraints). The pilot also provides feedback on timing to see if you need to spend more (or less) time on instructions and practice or on the main tasks. Last, you may learn that changes to payment procedures are required.

6.4 One Week Before: Almost There

One week before the experiment you should deal with the final preparations. At this point the devil lies in the detail. The most important check is of the practical and material details (instructions, lists, administrative papers) and managing subject subscription. The list of tasks is as follows:

1. Print the experimental instructions.
2. Print the consent and participation forms.
3. Prepare the experimental material if any (urns, calculators, pens, sheets, tables, etc.) and check the standard office supplies (printer cartridge, toner and paper).
4. Prepare the payment material. If you use cheques, be sure to have the right number of them. If you use cash, be sure to have enough change, with a sufficient number of

notes or coins. If you use electronic payments, be sure to have enough credit and fast verification procedures.
5. Send out the invitations to participants. The recruiting software allows you to avoid clashes with other lab bookings and experiments. Due to attrition, the over-recruitment of participants is recommended. It is common practice to recruit 10–25% more subjects than required. This precaution avoids no-show participants affecting the experiment. Ask subjects to arrive five to ten minutes before the scheduled start of the experiment.
6. Schedule the reminder. The reminder should usually be sent to registered participants 24–48 hours before the experiment.
7. Once the experimental sessions are complete, you can print the list of subjects for each session.
8. Prepare a script to test the program the day of the experiment. Make sure that you will be able to pay the subjects in case of a computer crash during the experiment.

6.5 D-Day: Step-by-Step Proceedings

6.5.1 Before the Participants Arrive

Before the first experimental session starts there are a number of checks to be made. To avoid rush and stress, you should take one hour to check that everything is in order for your first session. If the experiment is computerised, the computer (server/terminals) will take much of your time before the subjects arrive. Below you can find a tentative list of items that you should check before the participants arrive.

1. Start the computers/terminals and log on.
2. Check the language of the keyboards.
3. Clean the old data on the terminals.
4. Run the test script.
5. Check the exchange rate to be used during the experiment.
6. Launch the clients on the terminals.
7. Prepare the administrative tasks you will have to carry out after the experiment: list of payment receipts, list of consent forms, etc.
8. Put your mobile devices on vibrate mode.
9. Prepare the material needed for when the participants arrive: consent forms, list of registered participants and instructions.

6.5.2 Participants' Arrival

When participants arrive, you first have to check that their name appears on the convocation list, check their ID and have the participants sign the consent form (see Chapter 2, Figure 2.1, for an example). If the lab has a waiting room, you can invite the subjects to go there until the experiment starts. If over-recruitment produces more subjects than required for the experiment, you can propose that these subjects show up for another experimental session instead, or leave with the show-up fee.

When all subjects are ready, they enter the experimental lab. It is important for each participant to be randomly assigned a seat. One simple random-assignment method is to write the seat number on top of the instructions and distribute them randomly. Alternatively subjects can draw their seat number from an urn placed next to the entry door.

Now the experiment can start. You can read the instructions aloud or ask subjects to read the instructions on their own. Don't forget to ask subjects to switch off their own mobile devices. Figure 6.4 shows the main elements of a typical experimental session.

6.5.3 The End of the Experimental Session

At the end of the session, it is best to ask subjects to remain seated until they get paid. Print or fill out the payment receipts before proceeding to payment. Once subjects have been paid they can leave.

Once all subjects have left, you can prepare the room for the next session. Sessions do not follow each other immediately, to allow time to clean up the room, and to make sure that those leaving one session do not talk to those who are waiting outside for the next one to begin. This is also the time to deal with administrative tasks. First, you have to maintain the subject database, with the following steps:

- Check the number of subjects in the session. If the number of subjects is different from that scheduled, update the database with the right number of subjects. Update the show-up data for no-shows.
- Add new and unscheduled participants, if any. Check if they are correctly entered in the database. Update the registration details of the over-recruited subjects who chose to register in another session.
- For the subjects who showed up, update the show-up and participation data.
- Edit and close the session.

Second, you have to fill in the financial documents to track the money spent during the experiment and the receivers (amount, name, address, etc.). Third, save the experimental data to a safe device. Once this is done, the experimental session is over and you can prepare for the next session.

6.6 *Case Study*: Measuring Preferences in Choice over Time

Choice over time, which can be defined as any decision that requires trade-offs among outcomes that occur at different points in time, is of great importance in economics. The measurement of time preferences has received considerable attention in experimental economics and experimental psychology (Frederick et al., 2002) and a variety of methods and experimental procedures have emerged. In this case study, we review these methods and procedures in three steps. We first present simple methods to measure preferences in choice over time, as well as the main experimental challenges in this area, and then more sophisticated methods.

Table 6.1 Example of binary choices used by Tanaka et al. (2010)

Sooner–smaller	Larger–later
60 today	120 in 1 week
20 today	120 in 1 month
100 today	120 in 3 months
50 today	300 in 1 week
200 today	300 in 1 month
150 today	300 in 3 months
50 today	60 in 3 days
30 today	60 in 2 weeks
10 today	60 in 2 months

6.6.1 Simple Methods to Measure Preferences in Choice over Time

The experimental elicitation of time preferences is usually based on trade-offs between an earlier outcome (usually a smaller outcome, denoted SS, for 'smaller–sooner') and a later outcome (usually a larger outcome, denoted LL, for 'larger–later'). More complex choices involving temporal sequences of outcomes can also be used to elicit attitudes, but are less common (Rubinstein, 2003). Most elicitation methods used in decision over time are based on binary choices, indifferences or price lists. This subsection follows this simple classification and reviews the main elicitation methods available in the literature.

Binary Choices

A simple procedure that can be used to elicit time preferences is a series of binary choices. This procedure varies both the amount and delays in order to estimate a parametric discount function with one (exponential), two (quasi-hyperbolic, generalised hyperbolic) or even three or more parameters (fixed-cost discount function *à la* Benhabib et al. 2010). The number of binary choices used to elicit time preferences also varies from one study to another: for example Chabris et al. (2008) use 27 choices while Eckel et al. (2005) use 38 of them. The framing of the available options may also differ. A simple example of binary choices is provided by Tanaka et al. (2010). Table 6.1 shows nine choices out of the list of 75 they propose, each between a smaller–sooner outcome and a larger–later one.

Based on the answers to the binary choices, the researcher can appeal to discrete-choice models to estimate preferences. A basic estimation strategy works as follows. The researcher first specifies a decision model. Focus 6.1 describes the basic representation of the standard decision model in choice over time, the discounted utility model, and Focus 6.2 presents its behavioural foundations. In most of the experimental literature, the objects of choice are simple temporal prospects (x, t) denoting the receipt of outcome x at time t.[2] In the standard exponential discounting model of Samuelson (1937), the utility U of a temporal prospect (x, t) equals the present discounted value of the

[2] In most experiments no explicit distinction is made between payoffs and consumption; see Cubitt and Read (2007) for a discussion.

Focus 6.1
The discounted-utility model

Following Samuelson (1937) and Fishburn and Rubinstein (1982), a large part of the theoretical literature on choice over time appeals to discounted utility and additively separable functional forms. This discounted-utility model assumes a separation between value and delay in the evaluation of temporal sequences of outcomes. The discounted-utility model is the general behavioural theory for choice over time on which experiments are constructed. This model can account for most representations of choice over time, as well as the most common observed patterns of time preference. As such, the model is a very useful basis for experimental investigation. In the standard exponential discounting model of Samuelson (1937), the utility derived from a temporal prospect equals the sum of the present discounted values associated with each timed outcome. For example, for a temporal prospect yielding x at time t and y at time $t + \tau$, overall utility U is given by the sum of the present discounted values of x and y:

$$U(x, t; y, t + \tau) = \delta^t u(x) + \delta^{t+\tau} u(y) \tag{6.1}$$

where δ is the discount factor and u is a real-valued instantaneous utility function.

temporal prospect: $U(x, t) = \delta^t u(x)$, with δ the exponential discount factor. Most empirical work on time preferences assumes linear intertemporal utility. In this case, a unique preference parameter (δ) is estimated. In a series of choices $j = 1, \ldots, J$ between an immediate reward $(x_j, 0)$ and a delayed reward (y_j, t_j), the exponential discounting model predicts that the delayed reward is chosen if $U(y_j, t_j) = \delta^{t_j} y_j$ exceeds $U(x_j, 0) = x_j$. If we assume that the decision-maker makes this choice with some error, the decision becomes stochastic. A simple way of introducing error is to assume that it is identically and independently distributed over the J choices and follows an extreme-value distribution. In this case, the probability that the decision-maker chooses the delayed reward for choice j is

$$Pr[(y_j, t_j)|\delta] = \frac{exp(\delta^{t_j} y_j)}{exp(x_j) + exp(\delta^{t_j} y_j)} \tag{6.2}$$

This provides the likelihood contribution for a single subject's decision in choice j for parameter δ. For a given set of decisions $j = 1, \ldots, J$, the likelihood associated with a given subject's choices is

$$L(\delta) = \prod_{j=1}^{J} Pr[(y_j, t_j)|\delta]^{Z_j} \times [1 - Pr[(y_j, t_j)|\delta]]^{1-Z_j} \tag{6.3}$$

where $Z_j = 1$ if the subject chooses the delayed reward and 0 otherwise. Maximising this likelihood yields the estimated value of the preference parameter δ. In Chapter 7, Section 7.4, this basic econometric structure is extended to account for heterogeneity and a richer specification of the decision error.

Focus 6.2
Behavioural foundations of the discounted-utility model

Using simple temporal prospects, Fishburn and Rubinstein (1982) provide a behavioural foundation for canonical models of choice over time. They focus on a single outcome at a particular time (x, t), and assume that a preference relation is defined over the temporal prospects. They show that four axioms (order, monotonicity, continuity and impatience) are sufficient for the existence of a continuous utility function U representing the preference relation. When preferences satisfy these four axioms, the utility function is increasing in the outcome x, decreasing in the delay t for gains, and increasing in the delay for losses. *Order* is a rationality axiom defined in the usual fashion (the preference relation is reflexive, complete and transitive), while *continuity* is a mainly technical axiom. Under *monotonicity*, if an outcome x is preferred to an outcome x' in the present, it should also be preferred to x' when both outcomes are delayed to the same extent – if $x \succ x'$ then $(x, t) \succ (x', t)$. The last axiom, *impatience*, is based on two components. First, the decision maker is indifferent between receiving a zero outcome sooner or later. Second, desirable outcomes are preferred sooner rather than later. The axiom also states that the decision-maker procrastinates regarding losses: undesirable outcomes are preferred later rather than sooner. To obtain the standard exponential discounting model in (6.1), a fifth axiom is required. The *stationarity* axiom states that indifference (denoted $\sim$) between two temporal prospects depends only on the difference in delay between the two outcomes. If the decision-maker is indifferent between two temporal prospects, $(x, t) \sim (y, t + \tau)$, then the decision-maker will also be indifferent between these two temporal prospects when they are similarly delayed $(x, t') \sim (y, t' + \tau)$, whatever the values of t and t'. Under stationarity, a decision made at a given date does not change when the receipt periods are similarly delayed. The choice between two delayed outcomes then depends only on the time distance, τ, elapsed between them. Stationarity thus corresponds to constant impatience and underlies the evaluation of temporal prospects based on constant discount rates, as in the exponential-discounting model.

Indifference

Rather than using binary choices, it is also possible to elicit time preferences via indifference. Illustration 6.4 shows a set of written instructions used in such an elicitation experiment. This method offers a fairly simple and direct way of eliciting discount factors and discounting functions. For example, assuming a discounted utility model and a linear utility function, a single answer can be used to estimate a constant discount rate. Consider the following indifference between a smaller–sooner reward x at date t (denoted (x, t)) and a larger–later reward y at date $t + \tau$, (denoted $(y, t + \tau)$):

$$(x, t) \sim (y, t + \tau) \tag{6.4}$$

In the well-know contribution of Thaler (1981), y is elicited by asking subjects how much they would require to make waiting τ just as attractive as receiving x right now.

Illustration 6.4
Instructions for a time-preference-elicitation experiment

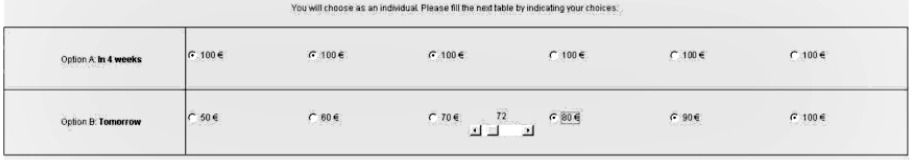
The following set of instructions is taken from Denant-Boèmont et al. (2017) (italicised text refers to the outline described in Section 6.2.1).

1. Welcome the participants: Thank you for participating in our experiment.

2. General description of the experiment: During this experiment, you will have to make decisions involving various amounts of money. If you follow the instructions, you could win a quite a large amount of money [*logic of incentives*]. All your responses will be converted into anonymous data after the experiment [*anonymity*]. During the experiment, you must answer a series of choice questions. There are no right or wrong answers to these questions. We are interested in your preferences: the only right answer to a choice task is the choice that you prefer [*absence of judgement*].

3. General structure of the experiment: Twenty people will participate in this experimental session. During the session, you will have to make decisions individually and collectively [*nature of interactions*]. Therefore, you will decide alone for some decisions and will interact with other participants for other decisions. For reasons of anonymity, you will not have access to the other participants' identities. The experiment consists of two parts [*timing of the experiment*]: in the rst part, you will decide as an individual; in the second part, you will make a decision in common as a member of a group of ve people (i.e. you and four other people). [*This part of the instructions will not be presented here*]

4. Description of the tasks: During the experiment, you will be asked to answer a series of choice questions regarding di erent amounts of money available on di erent dates. The gure below gives an example of one such series.



Option A o ers a xed amount of e 100 to be obtained in four weeks' time. Option B o ers a series of six amounts, equally spaced between e 50 and e 100, to be obtained tomorrow. For each of the six amounts, you will be asked to indicate whether you prefer option A or option B. Once you have switched between option A and option B, a scrollbar will appear on the screen. The scrollbar allows you to state the exact amount of money at which you switch your choice from A to B. For instance, imagine you decided to switch at e 72. If you switch at e 72, do you agree that you prefer to choose option B at a higher amount than e 72? (Y/N). Do you agree that you prefer to wait four weeks and choose option A at prices lower than e 72? (Y/N) [*pre-experiment questionnaire*]. If you have any questions, please feel free to ask the experimenter.

5. Description of the payment: Payments will be made as follows. At the end of each experimental session, four participants will be selected at random among the 20 participants in the session. For each of these participants, the computer will select one decision at random. For that decision, the computer will select one possible choice at random. Take the decision represented in the figure above as an example. For that decision, an integer between 50 and 100 will be selected at random. If the computer draws 63, then the selected choice is between €63 tomorrow and €100 in four weeks' time. Do you agree? (Y/N). If you chose €72 as the switching point, then your preference is for €100 in four weeks' time, and you will receive your payment directly by bank transfer from the National Treasury in four weeks' time. Do you agree? (Y/N). Suppose that instead the computer draws 83: your choice is then between €83 tomorrow and €100 in four weeks' time. Do you agree? (Y/N). If you chose €72 as the switching point, then your preference is €83 tomorrow and you will receive your payment directly by bank transfer from the National Treasury tomorrow. Do you agree? (Y/N). If you have any questions, please feel free to ask the experimenter.

6. End of the experiment: At the end of the experimental session, you will receive a receipt from the university as a proof of your payment.

Benhabib et al. (2010) uses the same design to elicit time preferences. Two treatments are implemented. In the first, subjects answer the following question:[3]

> *What amount of money,* $__*, if paid to you today, would make you indifferent to* $y *in* τ *days.*

The amount of the larger–later outcome y varies between \$10, \$20, \$30, \$50 and \$100, and the delays τ between three days, one week, two weeks, one month, three months and six months. The incentive scheme is a random-task Becker–DeGroot–Marschak procedure. First, one question (task) is picked at random. Second, a number is chosen at random in the interval $[0, y]$. If the number is less than x, then the subject receives y with delay τ; if it was greater than x, then the subject receives the payment corresponding to that number. Rather than directly asking subjects their indifference values, it is also possible to infer indifference values from a series of binary choices. Illustration 6.5 gives an example of such a ping-pong or bisection procedure.

Takeuchi (2011) also uses indifference to elicit time preferences. Instead of stating a smaller–sooner value that generates indifference to a larger–later reward, subjects have to state the longest acceptable delay τ after which receiving an amount of money y is as good as receiving x now:

> *To me, receiving* $x *today is as good as receiving* $y *in* __ *days.*

[3] In the second treatment, subjects are asked to state the amount of money y that would make them indifferent between x today and y in τ days, with x being based on the answers in the first treatment.

Illustration 6.5
Eliciting indifferences via bisection

The bisection consists of a series of binary choices, which are part of an iterative process focusing on subjects' indifference values. The table below illustrates the bisection method. At each iteration, subjects are faced with two temporal prospects, labelled A and B, where temporal prospect A always refers to an amount of money, x, now, and B refers to \$100 in three months' time. The chosen prospect is printed in bold. The starting value of the iteration for A here is the nominal value of B, although other values can be used. Depending on the choice made, the amount of outcome A rises or falls at the next iteration. In the table below, the size of this change is half the change in the previous iteration.

Iteration	Temporal prospect A	Temporal prospect B
1	**\$100 now**	\$100 in 3 months
2	\$50 now	**\$100 in 3 months**
3	\$75 now	**\$100 in 3 months**
4	\$87.5 now	**\$100 in 3 months**
5	**\$93.75 now**	\$100 in 3 months

Other step sizes are, of course, possible. After a given number of iterations (five here), the method produces an interval in which the indifference value should lie (here between \$87.5 and \$93.75 for a value of B of $y = \$100$ in three months' time). The indifference value is generally approximated by the mid-point of this interval. A more precise value can be obtained by additional iterations or the use of a scrollbar if the experiment is computerised.

where x and y are either \$5, \$10, \$15, \$20 or \$25, and $x < y$. Takeuchi (2011) also uses a Becker–DeGroot–Marschak incentive scheme: a number is picked from a uniform distribution between 0 and 120 days. If the number is less than the elicited delay τ, then the subject receives y with a delay corresponding to the number drawn; if it is greater than τ, then the subject receives the payment x at the end of the experiment.

We turn next to the elicitation of discount factors. Under the discounted utility model, the indifference in (6.4) can be represented as:

$$\delta(t) \cdot u(x) = \delta(t + \tau) \cdot u(y) \tag{6.5}$$

where $u(.)$ denotes the utility function and $\delta(.)$ the discounting function. Assuming a linear utility function, the indifference (6.5) yields:

$$\delta(t) \cdot x = \delta(t + \tau) \cdot y \tag{6.6}$$

Assuming constant discounting, the discount factor δ can be calculated directly and is $\delta = \left(\frac{x}{y}\right)^{1/\tau}$. In discrete time, the constant discount rate corresponding to this discount factor is $\delta^{-1} - 1$. It is worth noting that the elicited value of δ does not depend on the date at which the indifference is stated. This independence illustrates the implications of one

Focus 6.3
Accounting for non-linear utility

A great deal of work on choice over time assumes that utility is linear. For simple temporal prospects, this assumption is required for the identification of the discounting parameters (Fishburn and Rubinstein, 1982; Bleichrodt et al., 2009). If utility is non-linear, the imposition of linear utility will bias the measured discount factors $\delta(t)$ and discount rates. For example, if utility is concave, and assuming constant discounting, the discount factor δ calculated from indifference (6.4) is

$$\delta = \left[\frac{u(x)}{u(y)}\right]^{1/\tau}$$

This is higher than $\left(\frac{x}{y}\right)^{1/\tau}$, the discount factor from the assumption of linear utility. A number of methods avoiding the assumption of linear utility have been used in the literature. Abdellaoui et al. (2010) and Attema et al. (2010, 2016) measure discounting using methods that require no assumptions about utility. Andersen et al. (2008) and Takeuchi (2011) do not measure utility directly, but instead assume that utility in choice over time is equal to utility under risk, which is measured separately. This procedure was suggested by Frederick et al. (2002). One possible concern here is that measurements of utility under risk often assume expected utility theory to hold, which can distort measurement if it does not hold. Epper et al. (2011) elicit discount factors via indifferences similar to (6.4), together with a series of certainty equivalents under risk. The elicitation of certainty equivalents allows the estimation of the component of a non-expected utility model: utility and probability weighting. Epper et al. (2011) show that both probability weighting and the curvature of utility affect discount factors and discount rates. Moreover, departures from linear probability weighting are significantly correlated with departures from exponential discounting. This result shows that while utility curvature is important to avoid bias in the measurement of discount factor under hyperbolic discounting, probability weighting is the main channel by which uncertainty in future payments is linked to alternative models of discounting.

of the main axioms underlying the exponential discounting model, stationarity, whereby the choice between x and y depends only on the time distance τ between them. Here t might be the present or any delay, and this would not change elicited time preferences under stationarity. In equation (6.6), the discount factor is calculated under the assumption that utility is linear. Focus 6.3 discusses this assumption; in particular it shows how restrictive this assumption can be for elicited discount factors and discount rates, and how it can bias the elicited or estimated parameters.

Thaler (1981) considers a series of indifferences and shows that the elicited discount factor δ is not constant but increases with delay τ; discount factors also rise (so that the corresponding discount rates fall) as the size of the outcome x increases. Moreover, behaviour appears to be different in the loss domain.[4]

[4] Discount rates in the loss domain are found to be much lower than those in the gain domain, and changing τ has almost no effect on the elicited discount factors.

Table 6.2 The price list in Coller and Williams (1999)

Payoff alternative	Option A (in 1 month)	Option B (in 3 months)	Annual interest rate	Annual effective interest rate
1	\$500	\$501.67	2.00%	2.02%
2	\$500	\$502.51	3.00%	3.05%
3	\$500	\$503.34	4.00%	4.08%
4	\$500	\$504.18	5.00%	5.13%
5	\$500	\$506.29	7.50%	7.79%
6	\$500	\$508.40	10.00%	10.52%
7	\$500	\$510.52	12.50%	13.31%
8	\$500	\$512.65	15.00%	16.18%
9	\$500	\$514.79	17.50%	19.12%
10	\$500	\$516.94	20.00%	22.13%
11	\$500	\$521.27	25.00%	28.39%
12	\$500	\$530.02	35.00%	41.88%
13	\$500	\$543.42	50.00%	64.81%
14	\$500	\$566.50	75.00%	111.53%
15	\$500	\$590.54	100.00%	171.45%

Price Lists

A popular way of using binary choices in experimental economics is (multiple) price lists. A price list consists of a series of ordered binary choices. The Coller and Williams (1999) method is based on a bracketing of the indifference between two simple temporal prospects: a fixed smaller–sooner outcome and a varying larger–later one. The choices proposed by Coller and Williams (1999) are shown in Table 6.2. The prospect x is worth \$500 in $t = 1$ month,[5] and y varies between \$501.67 and \$590.54 in ascending order. The larger–later reward y is available in $t + \tau = 3$ months. The numbers from \$1.67 to \$90.54 are chosen to reflect discount rates from 2–100%. In each row, subjects pick their preferred option from A and B. Together with the description of the two options, Table 6.2 provides information on the annual interest rate and the annual effective interest rate.[6]

Coller and Williams (1999) use five treatments to address a number of possible experimental-design issues in choice over time – summarised in Table 6.3. Their baseline treatment does not provide any information on either market or experimental interest rates. The first four treatments consist of a 2×2 design, in which the first treatment variable is the information on market interest rates and the second treatment variable is information on experimental interest rates. Two additional experimental treatments test

[5] Coller and Williams (1999) choose the sooner delay t to be one month to minimise the perceived difference between option A and option B in terms of transaction costs and any uncertainty associated with future payments.

[6] Annual effective rates are calculated using daily compounding. Annual interest rates correspond to simple compounding. Coller and Williams (1999) also provide subjects with information on the market interest rate. They argue that the discount rates revealed in the lab are influenced by subjects' market opportunities. As these latter are mostly unobservable, the experimenter provides the subjects with some information about market opportunities.

Table 6.3 The treatments in Coller and Williams

Treatment	Front-end delay rates	Information on experimental discount rates	Information on market interest rates	Real payments	Results
1	X			X	0.221–0.284
2	X	X		X	0.162–0.191
3	X		X	X	0.191–0.221
4	X	X	X	X	0.162–0.191
5				X	0.284–0.419
6	X	X	X		0.105–0.133

Note. The results column shows the median interval of the annual discount rate.

the impact of the one-month front-end delay on outcome $x = \$500$ and the hypothetical bias in choice over time. The last column of Table 6.3 shows the interval of discount rates consistent with the median choices in Table 6.2.[7]

Coller and Williams (1999) show that provision of interest-rate information reduces the elicited discount rates and their variance. The use of a one-month front-end delay for the smaller–sooner outcome also reduces the elicited discount rates. The results in Table 6.3 are consistent with those in Kirby and Maraković (1995) that real incentives increase discount rates, but Coller and Williams (1999) mention this could be due to the fact that subjects were not perfectly randomised into treatments, with a possible correlation between demographics and the treatments. When the effect is controlled for, Coller and Williams (1999) obtain a marginally significant negative effect of real payment on discount rates.

The experimental design in Coller and Williams (1999) has been widely applied in the literature. For example, Andersen et al. (2008) reduce the list to 10 choices that are presented to field subjects in an artefactual field experiment. Option A offers 3,000 Danish krone in one month and option B 3,000 + x Danish krone in seven months, where x reflects annual rates of return of 5–50% on the principal of 3,000 Danish krone, compounded quarterly. The authors use six discount-rate tasks corresponding to six different time horizons: 1 month plus 1 month, 4 months, 6 months, 12 months, 18 months and 24 months. They also include four Holt and Laury (2002) multiple-price lists to elicit utility.

Manzini and Mariotti (2014) compare the price-list elicitation of time preferences with a Becker–DeGroot–Marshak mechanism based on matching. They find that the price-list method produces discount rates that are higher than those delivered by a BDM. The experiment also involves a payment mechanism via a Vickrey auction. In the latter, subjects are asked to state the minimum amount x they would accept at an earlier date rather than receiving a given amount later (y = 20 Euros or 50 Euros at τ = 1, 2 or 4

[7] This corresponds to the interval consistent with the median switching point, assuming linear utility and exponential discounting in continuous time. For example, in session 1 the median choice was to postpone payment for the first 11 choices, corresponding to a discount rate in the 22.1–28.4% interval. The median indifference value produces a discount rate of around 22.5%.

Table 6.4 Four procedures to elicit indifference in choice over time

	Value equivalence	Delay equivalence
Sooner–smaller	x	t
Larger–later	y	τ

months). The subject with the minimum amount wins, and receives the second-lowest amount immediately; subjects who lose instead receive y with delay τ. Manzini and Mariotti (2014) find no difference in the elicited discount rates in the matching procedure and the Vickrey auction. However, the price-list method yields higher discount rates than does the Vickrey auction.

6.6.2 Experimental Challenges in Measuring Time Preferences

Choosing an Experimental Design

Several experimental issues occur when it comes to the choice of an elicitation procedure. First, the experimental design can influence the elicited preferences. Manipulating the outcomes (x or y) or manipulating the delays – or dates – at which the outcomes are available (t or τ) has been found to change the elicited discount rates. Second, the response mode can also have an impact on elicited preferences. For example, the two well-known procedures to elicit indifferences, choice tasks and matching tasks, do not necessarily generate the same elicited discount rates.

The choice of variables to be varied in the experimental design can have a considerable impact on the elicited values and estimated preference parameters in two main dimensions. The first is the scale on which the indifference is elicited. As shown in (6.4), four quantities can be manipulated to determine indifferences. Two of these, x and y, are on the outcome scale. The other two, t and τ, are located on the time scale. Table 6.4 shows the researcher's four ways of eliciting indifferences in choice over time. In experimental economics, the elicitation of indifferences is mainly based on value-equivalence tasks rather than on delay-equivalence tasks. Among these, a large majority use smaller–sooner value equivalence methods, i.e. changes in x, to elicit indifference.

To look for framing effects in time-elicitation questions, Benhabib et al. (2010) applied two experimental treatments based on value-equivalence tasks. In the first, a smaller–soonner amount is elicited; in the second, implemented one day after the first, a larger–later amount is elicited. While violations of exponential discounting occur in both treatments, Benhabib et al. (2010) find that larger–later value-equivalence tasks are less prone to present bias than smaller–sooner value-equivalence tasks. Frederick et al. (2002) find that larger–later value-equivalence tasks generate discount rates that are dramatically higher than those in smaller–sooner value-equivalence questions. Delay-equivalence tasks are also subject to framing effects, such as temporal referencing (Frederick and Loewenstein, 2008). Last, it has been found that discount rates are higher when delays rather than dates are used to describe future payments (Read et al.,

2005; LeBoeuf, 2006). The second dimension is the date at which the options are available. In experiments on choice over time the earlier date does not need to be fixed at the present, and the smaller–sooner payment can be made, if possible, at any given date in the future. Some experiments exploit this possibility to introduce a 'front-end' delay to reduce overreactions to current outcomes. A front-end delay is a minimum amount of time between the present and the date at which the smaller–sooner option becomes available. As such, the experiment only involves future outcomes, which reduces any influence of the availability of current outcomes on decisions. Experimental evidence on the effectiveness of front-end delays is mixed (Cohen et al., 2016). For example, Coller and Williams (1999) find that a one-month front-end delay reduces discounting, while there is no effect in Holcomb and Nelson (1992). Augenblick et al. (2015) find small effects of front-end delays on monetary outcomes, but larger effects in a real-effort task.

The choice of the response mode is also of great importance for elicited preferences. A simple way of eliciting indifferences would be to directly ask subjects their indifference value – a procedure known as matching tasks. However, choices are generally considered more informative as they lead to fewer inconsistencies (Bostic et al., 1990). This is why procedures are mainly choice-based, in the form of choice lists, for example. In choice tasks, subjects indicate their preference over two (or more) options. In some experiments, subjects can also express indifference between the two options. Ahlbrecht and Weber (1997) compare choice and matching tasks in a treatment involving uncertainty over future outcomes in a within-subject experiment. In both tasks, the certain present value corresponding to a binary lottery is elicited. They find that while matching tasks produce decreasing impatience (a higher discount rate for short rather than long delays), choice tasks lead to fewer violations of the core axiom underlying the standard discounting model, stationarity. Overall, observed behaviour in the choice tasks appears to be much more in line with the standard exponential discounting model. In addition, Tversky et al. (1990) note that choice over time might be subject to preference reversals between choice and pricing (Lichtenstein and Slovic, 1971). Study 2 in Tversky et al. (1990) indeed shows that subjects tend to choose short-term over long-term prospects, even though they value the long-term prospect more highly. Read and Roelofsma (2003) specifically examine whether matching reveals more or less impatience than choice in a within-subject experiment, and find more impatience for choice than for matching. They also find supporting evidence for the result in Ahlbrecht and Weber (1997) that decreasing impatience is more likely to be observed in matching tasks than in choice tasks. To circumvent these difficulties, it is possible to elicit indifferences using a series of choices. A first way of doing so is via bisections – see Illustration 6.5 – which can be easier to answer for subjects than direct matching, as it only involves choices and does not require subjects to state a value for a temporal prospect. A second possibility is to use a series of ordered binary choices. This corresponds to the price-list method.

Choosing a Decision Model

Cohen et al. (2016) describe the main empirical regularities that have emerged in experiments with choices between smaller–sooner and larger–later outcomes. The review of the literature suggests six main regularities regarding the elicited discount rates:

1. Decreasing impatience: discount rates decrease over time.
2. Small effects of front-end delays: adding a fixed delay to both the smaller–sooner and larger–later outcomes barely changes the discount rates.
3. Sub-additivity: the elicited discount rates over longer time intervals are lower than would be expected from the elicited discount rates in a sequence of shorter sub-intervals.
4. Magnitude effect: discount rates fall with the size of the rewards.
5. Delay/speed-up asymmetry: discount rates are higher when the later outcome is described as delayed relative to the sooner outcome; discount rates are lower when the sooner outcome is described as advanced relative to the later outcome.
6. Gain/loss asymmetry: discount rates over gains are higher than those over losses.

Each of these regularities challenges the constant discounting-model that is usually assumed in the economics of choice over time. The shape of the discounting function continues to be debated, but a constant discount rate, which is stable across delays, is typically not supported by the data (Benzion et al., 1989; Thaler, 1981; Loewenstein and Prelec, 1992; Frederick et al., 2002). One major finding is that people are much more impatient when one choice option involves an immediate reward than when choices only involve future outcomes – a 'present bias'. For instance, discount rates in Thaler (1981) fall as the length of time to the larger–later payment rises. On the other hand, some recent work has shown that experimental controls for transaction costs and payment risk reduce present bias and produce results that are more consistent with constant discounting (Andreoni and Sprenger, 2012; Augenblick et al., 2015).[8]

Accounting for violations of constant impatience requires alternative decision models. These alternative models are particularly important to allow the researcher to measure the distance to the exponential discounting model. Fishburn and Rubinstein (1982) show how the discounted utility model can account for the observed violations of stationarity. Using a weaker condition than stationarity, known as the Thomsen condition, they show that an additive representation of preferences in choice over time exists in which the discount factor is not constant. In this case, the utility U derived from a temporal prospect (x, t) equals the present discounted value of the temporal prospect:

$$U(x,t) = \zeta(t)u(x) \tag{6.7}$$

where $\zeta(t)$ is the discount factor at date t. The flexibility of the discounted-utility model produces a variety of representations of preferences for choices over simple

[8] Attema et al. (2016) also find behaviour compatible with constant discounting.

prospects.[9] For tractability reasons, elicitation of time preferences often specifies a parametric function for $\zeta(t)$. We now present the main parametric forms for $\zeta(t)$ that have been considered in the literature.

Among the alternatives to the exponential discount-function (defined as $\zeta(t) = \delta^t$), one of the simplest is Herrnstein's (1981) proportional-discount function:

$$\zeta(t) = \frac{1}{1+\theta t}$$

with $\theta > 0$. Under proportional discounting, the shape of impatience decreases and impatience depends on both the evaluation date and the delay. Another alternative-discount function is the power discount function proposed by Harvey (1986):

$$\zeta(t) = \frac{1}{(1+t)^{\theta}}$$

with $\theta > 0$. A hyperbolic function can be approximated, in discrete time, by the popular quasi-hyperbolic discounting function of, e.g., Phelps and Pollak (1968) and Laibson (1997):

$$\zeta(t) = \beta\delta^t \text{ and } \zeta(0) = 1$$

where $0 < \beta \leq 1$ is the present-bias parameter and $\delta > 0$ is the discount factor. Assuming linear utility, the quasi-hyperbolic discount function is a genuine interpretation. It imposes a minimum penalty on all future rewards: the penalty is controlled by β, as any dollar available in the future is at most valued at $\$(1-\beta)$. This model has been quite popular in the literature because it is a convenient way of dealing with present bias (DellaVigna and Malmendier, 2006; DellaVigna, 2009; Laibson, 1997). First, the discount factor between two consecutive future outcomes is δ. The model therefore mimics the exponential discounting model as regards future consequences. Second, the discount factor between an outcome now and another at time 1 is different and is given by $\beta\delta$. The parameter β reflects the extra discounting associated with choices involving immediate outcomes. Instead of assuming that the present bias is a variable cost, Benhabib et al. (2010) introduce a fixed cost θ:

$$\zeta(t) = \delta^t - \frac{\theta}{y} \text{ and } \zeta(0) = 1$$

Loewenstein and Prelec (1992) assume that the discount function $\zeta()$ takes a generalised hyperbolic form:

$$\zeta(t) = \frac{1}{(1+\theta_a t)^{\frac{\theta_b}{\theta_a}}} \tag{6.8}$$

with $\theta_a, \theta_b > 0$. The parameter θ_a shows the difference from constant discounting (Rohde, 2010). The limiting case in which $\theta_a = 0$ corresponds to constant exponential discounting: $\zeta(t) = e^{-\theta_b t}$. The parameter θ_b can be interpreted as a discount factor.

[9] However, the model does have drawbacks. By summing the discounted values, the discounted-utility model assumes inter-temporal separability, which remains a questionable assumption despite its importance for the tractability of the model.

If $\theta_a > 0$, then discounting falls over time, and θ_a represents the distance from the exponential-discounting model.

Violations of stationarity can also be accommodated by more flexible non-hyperbolic discount functions. One limitation of hyperbolic discounting is that it cannot account for rising discount rates over time. By contrast, this is allowed by the constant absolute discount function of Bleichrodt et al. (2009), which is more flexible. The constant absolute decreasing impatience (CADI) discounting function is such that:

$$\zeta(t) = \begin{cases} e^{e^{-\theta t}-1} & \text{if } \theta > 0 \\ e^{-t} & \text{if } \theta = 0 \\ e^{1-e^{-\theta t}} & \text{if } \theta < 0 \end{cases} \tag{6.9}$$

Last, Ebert and Prelec (2007) and Bleichrodt et al. (2009) introduce the constant relative decreasing impatience discounting function (CRDI), as follows:

$$\zeta(t) = e^{-\theta_a t^{\theta_b}} \tag{6.10}$$

with $\theta_a, \theta_b > 0$. In contrast with typical discount functions, both the CADI and the CRDI can handle any degree of rising or falling impatience. These non-hyperbolic discount functions allow for increasing impatience, which has been found in some contributions (Onay and Öncüler, 2007; Attema et al., 2010; Takeuchi, 2011)

Choosing a Payment Procedure

In choice over time, the choice of payment conditions is a key element of the experiment. This points out the importance of the financial structure of a lab. Different procedures have been used to implement real incentives for temporal prospects. Andersen et al. (2008), Kuhn et al. (2014) and Denant-Boèmont et al. (2017) use treasury-based transactions to transfer the delayed payment into the subject's bank account. Anderhub et al. (2001b) and Coller and Williams (1999) provide a post-dated cheque to subjects. Chabris et al. (2008), Benhabib et al. (2010), Coble and Lusk (2010) and Abdellaoui et al. (2013) send a cheque to the subject's mailing address. Tanaka et al. (2010) appoint a village leader to deliver future rewards to subjects in the village. Takeuchi (2011) gives a US Postal money order. Andreoni and Sprenger (2012) use cheques sent to subjects' campus mailboxes, with a 100% on-time delivery promise from campus mail services. It is common practice to give subjects a show-up fee in addition to the experimental payment resulting from their choices. In choice over time, this show-up fee can be problematic as it concentrates the show-up fee in one period while decision-based incentives are spread over the possible delays offered in the experiment. To reduce the asymmetry between the show-up fee and the real incentives, Andreoni and Sprenger (2012) divide the former in two equally split payments: the first half to be paid at the time of the experiment, and the second half to be paid at a later date.

Another source of asymmetry between payments arises when payment on different dates generates different transaction costs (Holcomb and Nelson, 1992). In particular, if the present outcomes are paid in the lab at the end of a session, they might be disproportionately preferred by the subject when compared to future payments. The payment

procedure for delayed payment can increase transaction costs due to the time, effort and money subjects will have to spend to obtain their future rewards (coming back to the university to receive the reward, scheduling and meeting with the experimenter, or checking their mailbox). In practice, any transaction cost subtracts a premium from the reward. As present payments are free of transaction costs, they might be artificially overpriced by the subjects when delayed payments are subject to transactions costs (Kirby and Santiesteban, 2003). The easiest way to control for transaction costs is to equalise them between the available alternatives.

The use of a front-end delay (typically one day, one week or one month) allows the perfect replication of the payment procedure, and avoids any asymmetric treatment of the rewards due to transactions costs. Trust in the payment can be an issue since mistrust increases the risk associated with future payment. Experimental evidence shows that the presence of risk increases discount rates (Öncüler, 2000; Anderson and Mellor, 2009). Ahlbrecht and Weber (1997) find counterevidence to this conventional wisdom: in their matching tasks, subjects discounted certain outcomes more heavily than risky outcomes. This asymmetry disappears in choice tasks, where certain outcomes were not discounted more heavily. In economics, the use of a real incentive scheme acts as a strong constraint on the domain over which preferences in choice over time can be elicited. The use of a real incentive payment scheme places bounds on the possible delays. For example, Chabris et al. (2008) and Eckel et al. (2005) use delays between two and 186 days. Tanaka et al. (2010) vary delays from three days to three months. Andersen et al. (2008) vary delays between two months and 25 months. Andreoni and Sprenger (2012) vary delays between the present and 19 weeks. Most of the time, the real incentive scheme is carried out via a random-task incentive system, in which one choice made by the subjects is selected at random to be played out for real.

6.6.3 Sophisticated Methods to Measure Preferences in Choice over Time

Convex Time-Budget Sets

Andreoni and Sprenger (2012) propose the convex time-budget set method to measure preferences in choice over time via a series of variations in linear-budget constraints over earlier and later consumption. The use of simple temporal prospects raises an identification problem for the utility function, unless it is assumed to be linear. Andreoni and Sprenger (2012) underline that choices between simple temporal prospects abusively restrict the decision to corner solutions in the outcome space: choice is either a smaller–sooner or a larger–later prospect, without any possibility of mixing the two. The convex time budget allows just this kind of mixing to produce preferred interior solutions, if these exist. Decision-makers choose between two base prospects: one (x_b, t) offering x_b at date t and another $(y_b, t+\tau)$ with y_b at a later date $t+\tau$, or any compound prospect along the inter-temporal budget line combining these two prospects. In the (x, y) plane the intertemporal budget line is as follows:

$$y = y_b - \frac{y_b}{x_b} x \tag{6.11}$$

Table 6.5 The convex time-budget method

Decision	Payment	Alternative					
		1	2	3	4	5	6
1	Today	$19.00	$15.20	$11.40	$7.60	$3.80	$0
	and in 5 weeks	$0	$4.00	$8.00	$12.00	$16.00	$20.00
2	Today	$18.00	$14.40	$10.80	$7.20	$3.60	$0
	and in 5 weeks	$0	$4.00	$8.00	$12.00	$16.00	$20.00
3	Today	$17.00	$13.60	$10.20	$6.80	$3.40	0
	and in 5 weeks	$0	$4.00	$8.00	$12.00	$16.00	$20.00
4	Today	$16.00	$12.80	$9.60	$6.40	$3.20	$ 0
	and in 5 weeks	$0	$4.00	$8.00	$12.00	$16.00	$20.00
5	Today	$14.00	$11.20	$8.40	$5.60	$2.80	$ 0
	and in 5 weeks	$0	$4.00	$8.00	$12.00	$16.00	$20.00
6	Today	$11.00	$8.80	$6.60	$4.40	$2.20	$0
	and in 5 weeks	$0	$4.00	$8.00	$12.00	$16.00	$20.00

Note. For each decision, subjects indicate their preferences over the amounts to be received at date t: *today* and $t + \tau$: in five weeks' time by choosing one alternative in each decision. In Andreoni and Sprenger (2012), the delays were replaced by their calendar names and subjects had access to a diary. Instead of using monetary amounts, they offer individuals 100 experimental tokens to be allocated to the sooner or later payoffs in any integer proportion.

where $\frac{y_b}{x_b}$ is the gross interest rate between periods t and $t + \tau$. The elicitation procedure is based on variations in both the gross interest rate $\frac{y_b}{x_b}$ and the timing of payments t and $t + \tau$. Changes in the outcome scale, which correspond to changes in interest rates, identify the utility function's parameters; changes in the time scale, via both t and $t + \tau$, identify the discounting parameters. Andreoni and Sprenger (2012) explicitly assume a power utility function and a quasi-hyperbolic discount function.

A simple variant of the convex time-budget method, proposed by Andreoni et al. (2013), is shown in Table 6.5. In this example, subjects make six decisions. In each decision, subjects select a single alternative offering x now and y in five weeks. In Table 6.5, the six decisions are based on six different gross interest rates: y_b is kept fixed at \$20 while x_b varies to yield interest rates between 5.26% and 81.82%. The outcomes in alternatives 2 to 5 are linearly spaced values between 0 and either x_b or y_b.

Andreoni and Sprenger (2012) use the convex time-budget method with five alternatives and a 3×3 design with three sooner delays t (now, seven days and 35 days), and three later delays via τ (35, 70 and 98). Both t and τ were chosen to be multiples of seven to ensure that payments arrive on the same day of the week. Andreoni et al. (2013) choose a 2×2 design with two sooner delays t (now and 35 days) and two later delays via τ (35 and 63).

Andreoni et al. (2013) measure time preference via both the convex time-budget method and a so-called double multiple price-list method. In the latter, following Andersen et al. (2008), they estimate the curvature of the utility function using Holt and Laury's price list for choice under risk (also known as the 'double price-list method'). The results at the individual level reveal no significant correlation between the measures

of the curvature, the present bias and the discount rate across the two methods. This may be thought to be problematic for the internal validity of the use of experiments to measure preferences in choices over time. Andreoni et al. (2013) investigate the internal validity by performing both in-sample and out-of-sample prediction exercises. While both methods perform well in-sample, they yield strikingly different out-of-sample success rates. The out-of-sample performance of the convex time-budget method is similar to the in-sample performance of the price-list method, with an out-of-sample prediction rate of 86%.[10] The authors also use three indifferences to carry out strict out-of-sample comparisons between the two methods. For the first indifference, subjects expressed their willingness to accept a sooner payment in exchange for $25 as a later payment. The other two indifferences are hypothetical and ask about (i) the willingness to accept a sooner payment in exchange for $20 in a month, and (ii) the willingness to accept a later payment in exchange for $20 today. The convex time-budget set method correlates with the indifference values, with correlations ranging from 0.23 for the incentivised indifference value to 0.54 for the hypothetical indifferences. As a comparison, the price-list method is imperfectly correlated with the indifference measures, with correlation figures ranging from 0.08 for the incentivised indifference to 0.60 for the first hypothetical indifference. Last, both methods have limited predictive validity at the distributional level: neither method is able to replicate the distribution of the indifference values.

Direct Method

Both the convex time-budget and double price-list methods jointly estimate the utility and discount functions. Both methods avoid the potential problems associated with the assumption of linear utility. They also differ from each other in their identification strategies and noise specification (Andreoni et al., 2013). Attema et al. (2016) introduce an elicitation method based on choices over monetary flows that avoids the need to estimate a utility function. One advantage of this approach is that it allows the researcher to directly infer discount rates from choices over outcome flows without any additional costs in terms of econometric analysis or parametric fitting.

Attema et al. (2016) present subjects with profiles of payments over a time horizon of 52 weeks. A profile is a sequence of payments paying x_t at the end of week t. The experimental design is based on comparisons between two profiles. For example, subjects face a choice between a profile that offers an extra weekly payment of $20 in the next four weeks (from 1 to 4) or a profile with the same extra weekly payment in the nine weeks from 5 to 13. The discounted-utility model assumes that the value of a sequence is given by $\sum_{t=1}^{T} \delta(t)u(x_t)$. In our example, assuming that $u(0) = 0$, the choice is therefore:

$$\sum_{t=1}^{4} \delta(t)u(20) \text{ versus } \sum_{t=5}^{13} \delta(t)u(20)$$

As utility on both sides cancels out, the choice corresponds to a comparison of cumulative discount weights: $\sum_{t=1}^{4} \delta(t)$ versus $\sum_{t=5}^{13} \delta(t)$. Attema et al. (2016) use a series

[10] The in-sample performance of the price list is 89%. When applied to the data obtained with the convex time-budget method, the price-list individual estimates predict only 16% of the choices.

Table 6.6 The choice list in the direct-method elicitation

Which option do you prefer? Gain $20 per week	
Profile A	Profile B
in week 1	Starting week 1 and ending (after) week 13
in week 1	Starting week 2 and ending (after) week 13
Starting week 1 and ending (after) week 2	Starting week 3 and ending (after) week 13
Starting week 1 and ending (after) week 3	Starting week 4 and ending (after) week 13
Starting week 1 and ending (after) week 4	Starting week 5 and ending (after) week 13
Starting week 1 and ending (after) week 5	Starting week 6 and ending (after) week 13
Starting week 1 and ending (after) week 6	Starting week 7 and ending (after) week 13
Starting week 1 and ending (after) week 7	Starting week 8 and ending (after) week 13
Starting week 1 and ending (after) week 8	Starting week 9 and ending (after) week 13
Starting week 1 and ending (after) week 9	Starting week 10 and ending (after) week 13
Starting week 1 and ending (after) week 10	Starting week 11 and ending (after) week 13
Starting week 1 and ending (after) week 11	Starting week 12 and ending (after) week 13
Starting week 1 and ending (after) week 12	in week 13
Starting week 1 and ending (after) week 13	in week 13

of such choices to elicit the week j at which the subject is indifferent between the two profiles. Table 6.6 shows a price list used in this experiment to elicit indifference.

If the subject switches from profile B to profile A between rows 6 and 7 (i.e. between five and six weeks), then the indifference value is 5.5. This means that the subject is indifferent between receiving $20 per week during weeks 1 to 5.5 and the same payment during weeks 5.5 to 13. Thus 5.5 weeks is the subjective midpoint of the time interval between the present and week 13. This subjective midpoint 'cuts' in two parts the cumulative discount weight between now and week 13. A patient enough subject will switch from profile B to profile A between six and seven weeks; an impatient subject will switch earlier. Lower values of the elicited subjective mid-point (in terms of weeks) thus correspond to greater impatience.

The general principle of the direct method is to elicit subjective mid-points of time intervals from indifferences, and to use these to measure a cumulative weighting function. Attema et al. (2016) show how the cumulative weighting function can be measured with precision by a series of indifferences. The first step is to elicit the time point $t_{.5}$ such that the subject is indifferent between receiving $20 per week during weeks 1 to $t_{.5}$ and the same payment during weeks $t_{.5} - T$, with end period T. Again, $t_{.5}$ 'cuts' the cumulative discount weight between the present and end week T in two. Two additional time points $t_{.25}$ and $t_{.75}$ are then established. The former splits in two parts the cumulative discount weight between the present and week $t_{.5}$, and the latter does the same for week $t_{.5}$ and T weeks. These subjective mid-points can thus be translated into discount factors $\delta(t)$ and used to estimate parametric discount functions.[11] The experimental design also elicits discount factors in a standard way, using indifferences between smaller–sooner and larger–later payments. The results in Attema et al. (2016) reveal less discounting

[11] For example, the discount factor $\delta(t)$ is the average of the derivative of the cumulative discount weight over the interval $(t-1, t)$.

under the direct method than with the standard method, even after correction for the curvature of the utility. In addition, the discount factors in the direct method were more in line with exponential discounting than were those from the standard method.

The Time Trade-Off Method

For discrete outcomes, Attema et al. (2010) introduce the time trade-off method as a way of sidestepping utility in the measurement of discounting. As with the direct method, this allows the direct measurement of impatience without any additional parametric assumptions regarding the discounting function. The time trade-off method elicits a sequence of n points in time based on a series of indifferences. The outcomes are kept fixed during this elicitation process. Indifferences are elicited by matching. For the first indifference value, subjects are asked to determine τ_1 such that receiving \$700 today (τ_0) is as good as receiving \$900 in τ_1 months. For the second, subjects are asked to determine τ_2 such that receiving \$700 in τ_1 months is as good as receiving \$900 in τ_2 months. The subsequent elements of the sequence $\tau_3, \ldots, \tau_n$ are elicited in the same fashion. In the sequence, the delay between the two consecutive dates exactly offsets the gain in the outcome. The delay between now and τ_1 months offsets the \$200 greater income (for example from \$700 to \$900), as does the delay between τ_1 and τ_2 months. The sequence measures the subject's willingness to wait to receive \$200 more at different points in time. The willingness to wait is the difference between two successive elements $\tau_j - \tau_{j-1}$ in the time trade-off sequence.

The time trade-off sequence also yields information on impatience. If the time trade-off sequence $(\tau_0, \tau_1, \ldots, \tau_n)$ reveals a constant willingness to wait, then subjects have constant impatience. In axiomatic terms, this means the stationarity axiom holds. If the sequence of elicited values $(\tau_0, \tau_1, \ldots, \tau_n)$ reveals an increasing willingness to wait, then the subjects have decreasing impatience. Moreover, subjects with a higher willingness to wait at the beginning of the sequence, when τ_0 is the present, have present-biased preferences. If the sequence has decreasing willingness to wait, then the subjects exhibit increasing impatience as time progresses. The major advantage of the time trade-off method is to characterise the discounting pattern without any assumption regarding the discount function or the need to elicit the utility function. The experimental results in Attema et al. (2010) show that subjects are increasingly impatient in the beginning and near future, and constantly impatient for points later in time. A graphical representation of the time trade-off sequence $(\tau_0, \tau_1, \ldots, \tau_n)$ with a time trade-off curve appears in Figure 6.5. The time trade-off curve plots the logarithm of the discount rate as a function of the point in time at which trade-offs are made.[12] A constant slope reflects constant impatience, and a convex (concave) time trade-off curve indicates decreasing (increasing) impatience. Figure 6.5 depicts the typical pattern found in the literature: decreasing impatience for the near future and then constant impatience. Bleichrodt et al. (2016) use this method to compare discounting for health and money, and find

[12] The logarithm of the discount rate, when normalised at τ_0 and τ_n, is $1 - \frac{j}{n}$ with $j = 1, .., n$. In other words, for each element τ_j on the x-axis, the time trade-off curve maps to the value $1 - \frac{j}{n}$ on the y-axis.

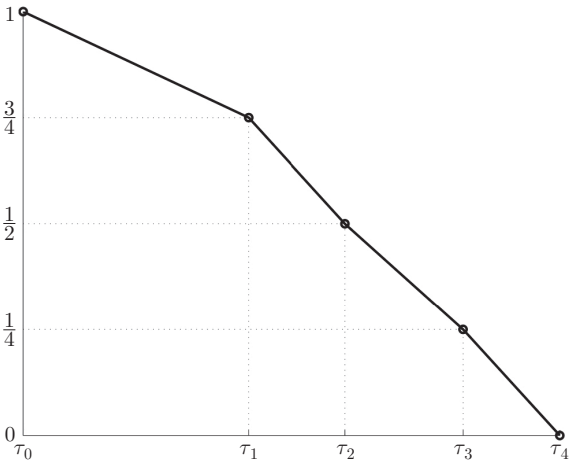


Figure 6.5 An example of a time trade-off curve

that departures from constant discounting are more pronounced for health than for money.

Summary

Conducting an experiment involves a series of challenges. The biggest is certainly building a laboratory, which requires important investment in money, time and organisational effort. Indeed, a lab is not only one or two rooms filled with computers and dedicated to running lab experiments. A series of computer infrastructure management, recruitment and financial procedures also has to be set in order to have an efficient operating structure. In this chapter, the building blocks of an experiment are (somewhat arbitrarily) organised along a timeline of the issues to be solved before the experiment starts. Aside from the existence of a lab, two elements are crucial for an experiment. The first one is the protocol. The second is the instructions. For both newbies and experienced experimenters, running several pilots is a perfect way to test an experimental design. Pilots are central to check that both the protocol and the instructions are consistent and that they fit the research objectives. Piloting is also a way to improve the experimental design, based on the feedback it generates. Last, in experimental economics, designing incentives is key and is an important element in the practice of experiments. The design of incentives can be easy for simple experiments. It can also become quite complex when experimental outcomes become more subtle, as in the case of choice over time.

7 The Econometrics of Experimental Data

Experiments allow us to create simple, controlled and incentivised environments in which economic agents make decisions. The first two parts of this book have underlined the need for experiments in economics. We saw that experiments have two main features that make them attractive. First, experiments allow the researcher to observe and measure parameters that would otherwise be unobservable. A typical example is the reservation price in an auction. Second, experiments provide control over the environment in which people behave – in such a way that the researcher is the one who decides on the data-generating process.

In most experiments, the parameters of interest belong to one of three categories. The first are behavioural parameters, which describe behaviour in a controlled environment, and their determinants. Such observed values can be compared to theoretical predictions (e.g. the observed level of contribution in a public-good game and the associated predicted Nash equilibrium) and potential explanations for any differences (e.g. can reciprocity and/or altruism explain the observed deviations from the Nash equilibrium?). The second category consists of the comparison of parameters between different experimental conditions, in line with exercises in comparative statics. In experiments, comparative statics are measured by treatment effects. In a public-good game, for example, the question whether additional information about the behaviour, or the rank, of others changes contributions would fall in this category. The last category is related to heterogeneity in observed behaviour. Typical examples are observable heterogeneity via gender effects, or the role of unobservable factors.

This chapter provides an overview of the econometric techniques that allow the researcher to tackle these three kinds of empirical question. It first sets out what experimental data are, and then discusses a set of methods for exploratory analysis. It then turns to empirical analysis, including both estimation and testing procedures, both parametric and non-parametric. Last, several specific parametric econometric models are presented, together with some applications to experimental data. The chapter ends with a brief description of more advanced econometric models. The extensive review of the econometrics of experimental data provided by Moffatt (2015) is recommended for use as a companion to this chapter.

Table 7.1 An example of experimental data based on second-price auctions

		Participant	Round	Group	Resale value	Individual bid	Highest bid	Second-highest bid
round 1	group 1	1	1	1	38	36	42	36
		2	1	1	43	42	42	36
		⋮	⋮	⋮	⋮	⋮	⋮	⋮
		9	1	1	32	31	42	36
	group 2	1	1	2	18	16	22	16
		⋮	⋮	⋮	⋮	⋮	⋮	⋮
		9	1	2	2	22	22	16
round 2	group 1	1	2	1	14	13	13	12
		2	2	1	12	12	13	12
		⋮	⋮	⋮	⋮	⋮	⋮	⋮
	group 2	1	2	2	41	41	80	56
		⋮	⋮	⋮	⋮	⋮	⋮	⋮

7.1 Experimental Data

Our working example is the experimental data delivered by second-price auctions, as presented in Chapter 2. Consider an experiment consisting of a market with one single seller and $n = 9$ buyers. The experiment has 18 participants, who are randomly matched at the beginning of the experiment into one of two groups of nine participants. Only one unit of the good is to be sold on that market. Each buyer is first shown a resale value on the screen and then submits a bid to buy the unit of the good. When all buyers have submitted their bid, the computer determines the ranks of all bids. The winner of the auction is the buyer whose bid is ranked first. The second-highest bid determines the market price – what the buyer needs to pay. The profit for the winner of the auction equals the difference between the resale value and the market price. All other buyers make zero profit. Once the subjects have performed the task, they move on to a new choice round.

7.1.1 A Working Example

Table 7.1 shows the data obtained from the experiment. For each participant and each round, the data include the group to which a participant belongs (third column), the resale value (fourth column), the bid (fifth column) and the results of the auction: the highest bid (which determines the winner) and the second-highest bid (which determines the market price). There are in addition a number of other variables in the data set (not shown in the table): the session date, the location of the experiment, the number of subjects in a given session, the starting time of the experiment, the duration of the experiment and/or the task, the name of the person(s) who ran the experiment, the payment made to subjects, the form of payment (cash, bank transfer, cheque) and so on.

The data as whole are of the panel type: the same participants are observed at different times (rounds) in the experiment. At a given round (e.g. round 1), the data in Table 7.1

is a cross-section: different subjects are observed at the same time in the experiment. Within a given round, the comparison across subjects documents individual heterogeneity. On the contrary, focusing on one participant (or a group average) over different rounds in Table 7.1 produces time-series data that allow us to measure changes over time in economic decisions. For example, if the list of induced resale values is kept constant during the experiment and simply rematched over buyers, the time series of the group average reveals the effect of learning and experience over the course of the experiment.

Experimental data are expected to be of higher quality than typical economic survey data. Missing answers or incomplete responses are relatively rare in experimental data sets. In the example in Table 7.1 there are no missing answers or incomplete responses. Every participant undertook the task in each round. The control that the researcher has over the experimental design also reduces measurement error and noise.

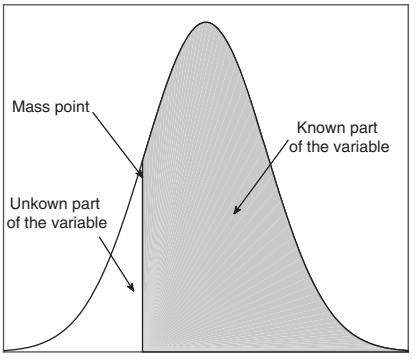
7.1.2 Types of Data and Measurement Scales

Before carrying out statistical analysis, the variables of interest have to be measured. Measurement depends on the type of variable in the experiment. We usually distinguish four measurement scales (also called types of data) in order to assign a real number to each element in the set of observations: the nominal, ordinal, interval and ratio scales. These different measurement scales require different statistical and econometric techniques. For example, the central tendency can be represented by the median or the mode with an ordinal scale, but not by the mean. On the contrary, the median, the mode and the mean can all be used for interval or ratio scales.

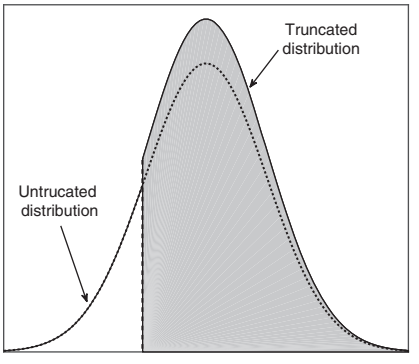
- A **nominal scale** is used for unordered categorical variables (like colours). A nominal scale with only two categories is called dichotomous, and corresponds to a binary variable. Typical examples of nominal scales are yes/no answers, gender, hair colour and location. Nominal scales are mutually exclusive, are not rankable, and basically represent different kind of things or people. In Table 7.1 the participant and group numbers are nominal scales.
- An **ordinal scale** is used for ordered categorical variables. This scale is used to order the variables by rank, although the size of the differences between the variables is unknown. Ordinal scales allow monotone increasing transformations of the variables. Typical examples of ordinal scales are rankings and non-numerical judgements or concepts such as the degree of agreement or disagreement, satisfaction or discomfort. From Table 7.1, the rank of the bid in the list of all bids in a group (first, second, . . . , last) is measured on an ordinal scale.
- An **interval scale** is a numeric scale in which differences between the numbers reflect known differences in the attribute. Any affine transformation can be applied to interval scales: the origin (the intercept) and the unit of measurement (the slope) are arbitrary. The typical example of an interval scale is temperature in Celsius or Fahrenheit. Interval scales are practical as the increments between values are known and measurable.

Focus 7.1
Censored and truncated data

Experimental data are sometimes censored or truncated. Censoring occurs when the observed value of some variable is only partially known. For example, in an experiment we might only measure buying prices in an auction up to a certain maximum price. Participants who report this maximum thus have a buying price that is at least the maximum price, and may well be higher. Truncation is different, as the elicited values can never be outside a given range or interval. For example, in experiments, prices are truncated at 0. In that sense, truncation is a characteristic of the distribution from which the data are drawn, while censoring is a (default) characteristic of the sampling procedure itself.



(a) Censored data



(b) Truncated data

The figure above shows the difference between censored and truncated data. In panel (a), censoring generates a mass point at the threshold where it occurs. Without censoring, the distribution would be the same, and the observed values above the threshold would be identical. For all values below the threshold, the variable takes the threshold value. Panel (b) shows that truncating, by contrast, modifies the density and does not generate a mass point at the threshold value. Were the truncation to be removed, the density, and therefore the observed values, would be different.

- A **ratio scale** is a numeric scale for which the differences and ratios between the numbers reflect differences and ratios in the attribute. Only affine transformations with a zero intercept can be applied to ratio scales: only the unit of measurement (the slope) is arbitrary and zero has a meaning. Typical examples of variables measured on a ratio scale are duration, length and prices.

7.1.3 Sampling

Most statistical methods assume that the samples are randomly selected from the population. If the samples are, however, not random, the typical statistical properties of the estimators are not satisfied and inference will be challenged – see Focus 7.1 for an example based on truncation and censoring. In practice, most samples are not entirely random and analysis is carried out *as if* the sample was selected at random. A random sample

is such that each possible sample of the same size has the same probability of being selected. Here, all individuals are equally likely to be sampled: if we number all individuals in a population from 1 to N, we then select n values randomly without replacement. If the sample size n is small relative to the population size N, then sampling without replacement will not much affect the assumption of independence between observations (see Chapter 3, Section 3.1.2, for a discussion). However, with a sample size n that is large compared to N, sampling without replacement will produce dependent observations. One limitation of random samples is that they may not allow the statistical analysis of small subgroups in the population.

In experimental economics, those samples that are easily available to the researcher are unlikely to be random. The very reason why samples are easily available (e.g. volunteers to participate in an experiment at a given site) might give rise to selection and affect the quality of the statistical analysis (see Chapter 8, Section 8.3.3, for a detailed discussion). Some alternatives to random samples are stratified random samples and clustered samples. In a stratified random sample, the population is divided into a number of homogeneous, non-overlapping groups (strata) with respect to some characteristics. Random samples are then taken within each stratum. The number of subjects per stratum is proportional to the stratum size in the population. A variance correction within each stratum can be applied. In this case, the sample is based on Neyman's optimal allocation, a method to maximise survey precision for a given sample size.[1] Cluster sampling is two-tier in nature. The population is first divided into a number of similar-looking heterogeneous groups. One group is then sampled at random and, within this selected group, individuals are again sampled at random. As the groups are similar to each other, the data from one cluster is representative of the whole population.

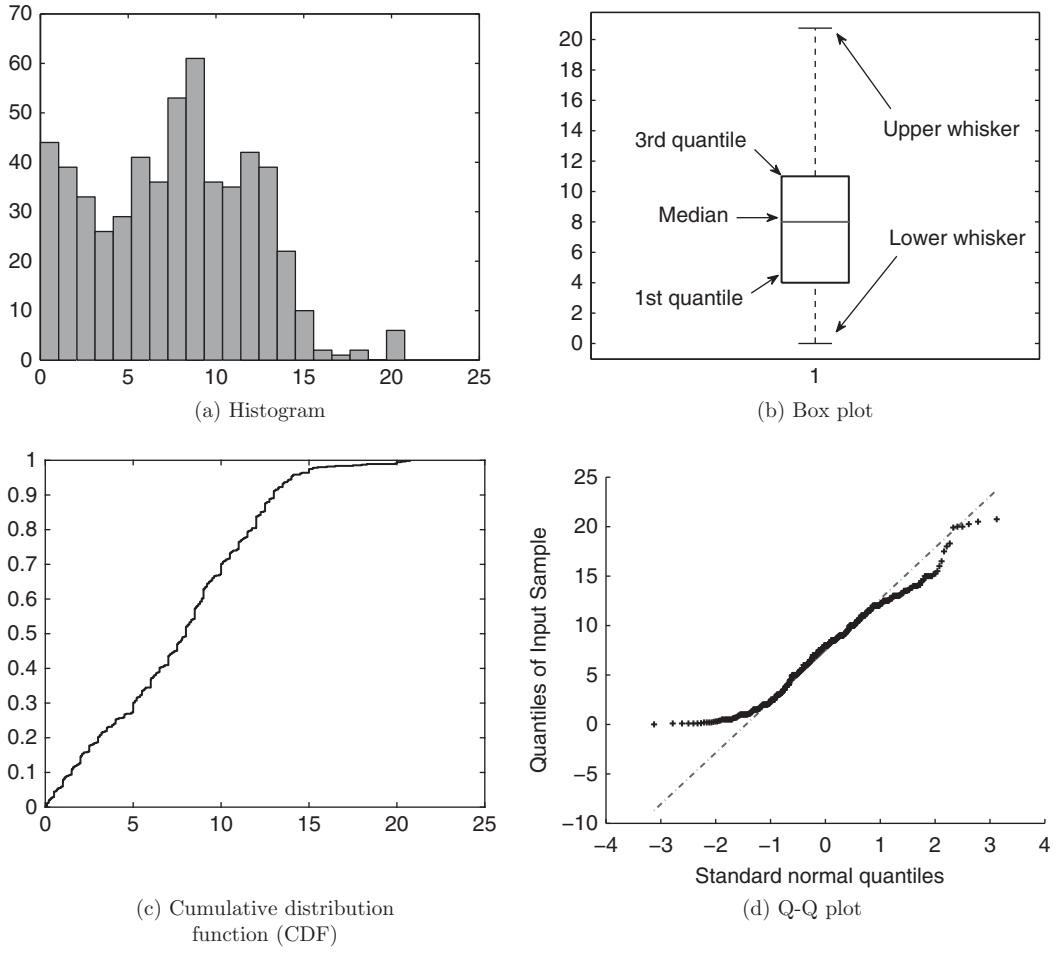
7.1.4 Exploratory Analysis

Exploratory analysis is the first step of the empirical analysis and aims to summarise the data before performing the statistical analysis. The summary produced will depend on the measurement scale of the variable. For categorical variables, this is limited to the range and mode of the variable, and the frequency of each value. For other quantitative data, the summary is broader and includes the centre of the distribution (the median and mean, if any), the spread (percentile, interquartile range, range, variance, if any), modality (i.e. the number of peaks in the density), the shape (i.e. the heaviness of the tails) and the number of outliers. Together with the descriptive statistics, the data can be efficiently summarised by a number of simple graphs.

Graphical Tools

Figure 7.1 shows four simple plots describing the same data: a histogram, a box plot, the cumulative distribution function (CDF) and the Q-Q plot.

[1] For a given stratum h, the sample size for this stratum is equal to the fraction $N_h s_h / \sum_k (N_k s_k)$ of the total sample size, where N_k is the population size for stratum k, and s_k is the standard deviation of stratum k.



(a) Histogram

(b) Box plot

(c) Cumulative distribution function (CDF)

(d) Q-Q plot

Figure 7.1 Visual representations of data

Histogram

This bar chart shows the distribution of the variable, with each bar representing the proportion of the observations that lie within a given range of values. A histogram is a simple way of seeing if the data are symmetric or skewed, unimodal or multimodal, hump-shaped or U-shaped, etc.

Box plot

This is a visual presentation of information about the central tendency, symmetry, skewness and outliers. The central mark is the median, the edges of the box are the first and third quartiles, and the whiskers extend to the most extreme data points that are not considered to be outliers. The actual outliers, if any, are plotted individually. Outliers are usually defined as data points that are more than 1.5 interquartile ranges distant from the closest interquartile range boundary. When the data are symmetric, the median is

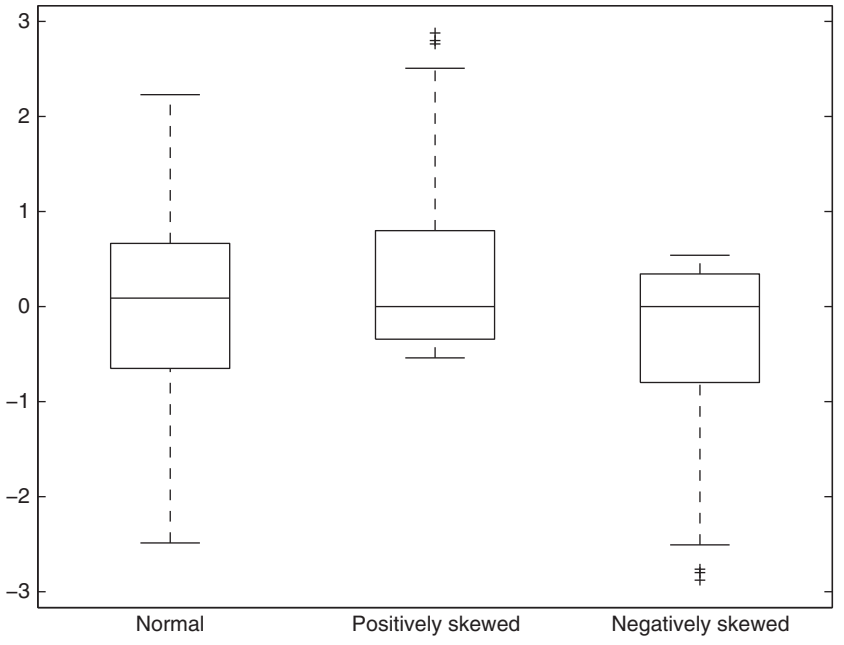


Figure 7.2 Box plots for different distributions

in the middle of the box and the whiskers are of the same length. A skewed distribution will have the median closer to the shorter whisker. If the top whisker is longer, then the distribution is positively (i.e. right-) skewed. If the lower whisker is longer, then the distribution is negatively (i.e. left-) skewed. As opposed to the histogram or the normal probability plot, the box plot does not reveal the presence of any multimodality. Figure 7.2 shows the box plots for normal and positively and negatively skewed distributions. The points outside the whiskers are the outliers.

The empirical cumulative distribution function

This is an ordered statistic of the sample, in which the observations are ranked from the lowest to the highest. The empirical cumulative distribution function shows the proportion of the sample for which the value of a particular variable is below any given number. It thus delivers the probability of drawing a value lower than that given number in the sample. The slope of the cumulative distribution function (CDF) is the density. In Figure 7.1, the comparison of panels (a) and (c) reveals that higher-density areas in the histogram (on the left) have steeper slopes in the CDF (and lower-density areas have flatter slopes). We can compare the empirical CDF to a reference distribution. For example, the CDF of a uniform distribution is a straight diagonal line. The empirical CDF is an effective way of detecting outliers, which are located at the tails of the distribution.

Normal probability plot

This plots the normal sample statistics against the quantiles of a standard normal distribution. The further the points are from the diagonal, the greater the departure from

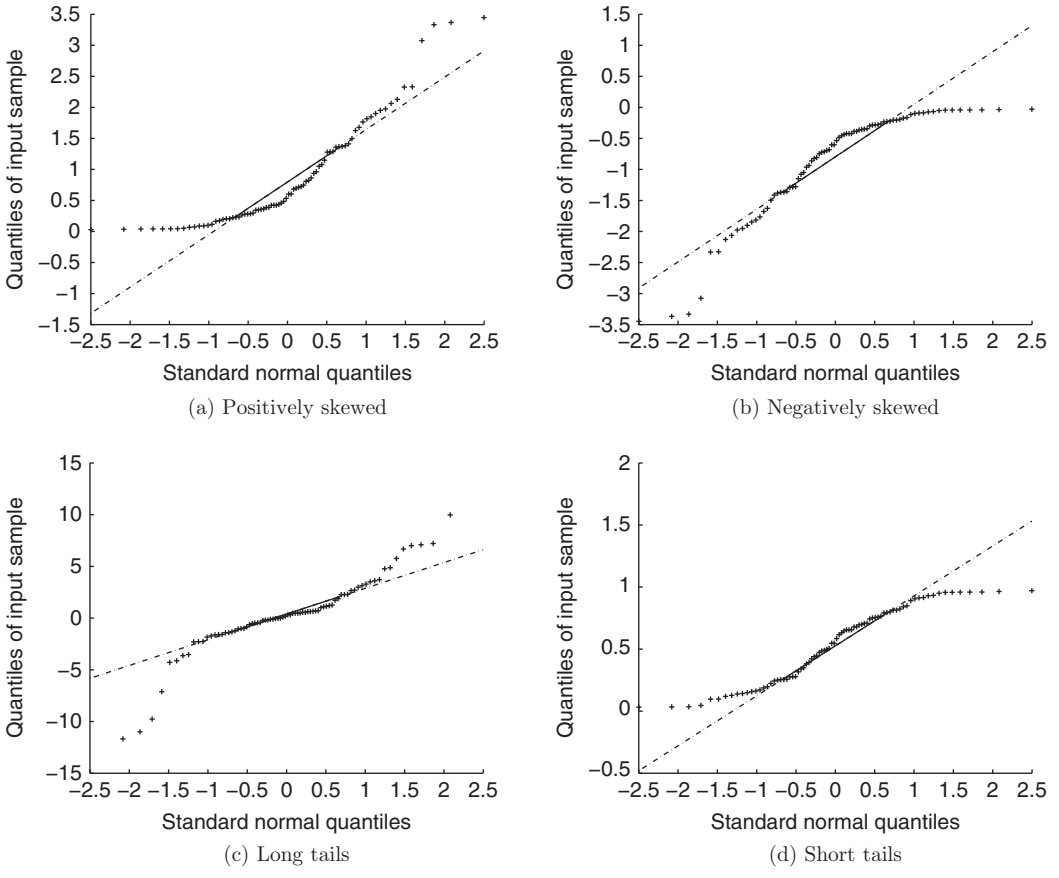


Figure 7.3 Normal probability plots

normality. The shape of the plot indicates the type of departure from normality. Figure 7.3 shows four different cases. In panel (a) the U-shape indicates a positively skewed distribution, and in panel (b) the hump-shape reflects a negatively skewed distribution. The normal probability plot has an S-shape in both panels (c) and (d). In panel (c), the points to the left are below the line and those to the right are above the line: this indicates long tails. In panel (d), the points to the left are above the line and those to the right are below the line: this indicates short tails. Other features of the normal probability plots are also of interest. Points located far from the main scatter plot correspond to outliers. If two separate scatter plots show up, the distribution is bimodal. The normal probability plot is a sub-case of the more general Q-Q plots that plot the quantiles of two distributions against each other. If the two distributions are identical, the Q-Q plot is the diagonal straight line. If the Q-Q plot is any other straight line, then one distribution is a linear transformation of the other and they differ in their location parameter (e.g. their mean, their median or their mode). If the Q-Q plot is steeper than the diagonal, the distribution plotted on the y-axis has greater dispersion than that on the x-axis. The

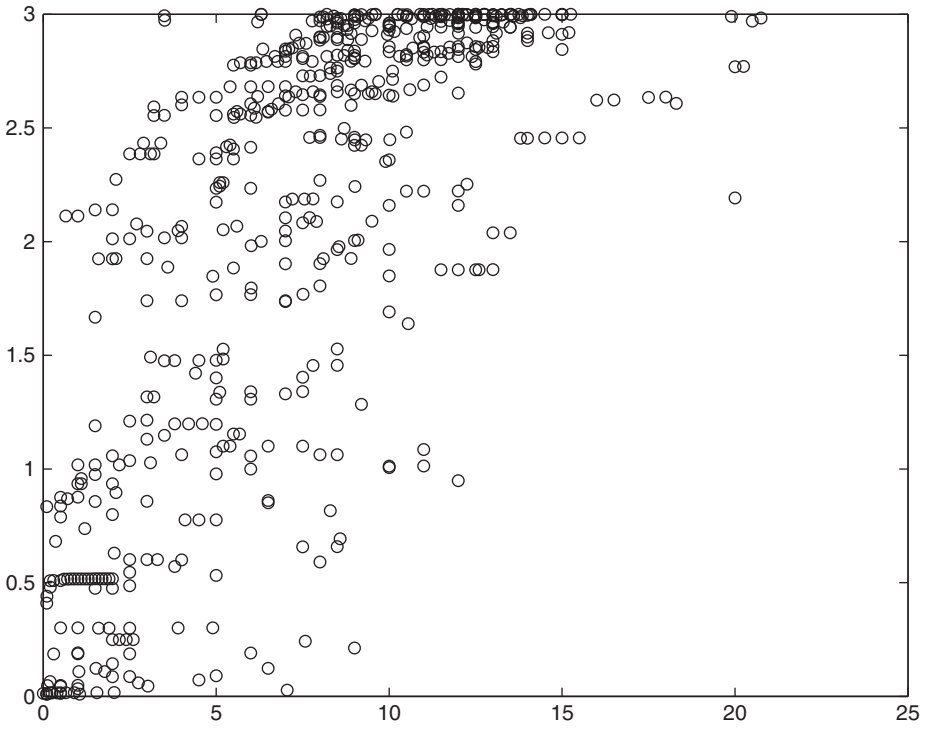


Figure 7.4 A scatter plot

intercept and the slope of the line in a Q-Q plot reveal the different location and scale parameters of the two distributions.

Scatter plot

This is the most useful graph when there are only two variables. The data are displayed as a collection of points, with one variable on the horizontal axis and the other on the vertical axis. Figure 7.4 provides an example. This plot is well suited to the exploratory analysis of the degree of association between two variables. In experiments with repeated measures, a scatter plot displaying one variable plotted against time is called a run chart.

Descriptive Statistics

The mean and the median are two measures of central tendency with different properties. The mean is sensitive to all data points, while the median is less influenced by outliers and skewed distributions. The mean cannot be used for ordinal variables. There are a number of other measures of central tendency. For example, the trimmed mean is the mean after discarding a given (generally equal) part of the distribution at the top and bottom. Dropping the lowest and highest 25% produces the 25% trimmed mean, also known as the interquartile mean. Another alternative is the Hodges–Lehmann estimator

Table 7.2 Descriptive statistics

Measure	Name	Value with outliers	Value with outliers
Central tendency	Mean	7.62	7.82
	Median	8.00	8.00
	Interquartile mean	7.73	7.91
	Hodges–Lehmann	7.55	7.75
Dispersion	IQR	7.00	6.60
	Standard deviation	4.35	4.41
	Coefficient of variation	0.57	0.56
	IQR/median	0.87	0.82
	Median absolute deviation	4.67	4.82
	Mean absolute deviation	3.58	3.61

Note. Descriptive statistics computed without (first column) and with (last column) the top and bottom 1% of the data displayed in Figure 7.1.

of central tendency, corresponding to the median of all pairwise means. Table 7.2 reports a variety of central-tendency measures for the data depicted in Figure 7.1, both with and without the top and bottom 1%.

Two common measures of dispersion, with different properties, are the standard deviation and the interquartile range (IQR). The IQR is not affected by extreme values and is less affected by skewed data than is the standard deviation. One difficulty with the IQR is that it cannot be easily mathematically manipulated (derived, for example). The standard deviation is a useful indicator of dispersion for distributions that are roughly normal. For a normally distributed variable, around 68% of observations lie between the mean and $\pm$ 1 standard deviation, 95% of observations between the mean and $\pm$ 2 standard deviations, and 99.7% of observations between the mean and $\pm$ 3 standard deviations. For ratio scales, the coefficient of variation (the ratio of the standard deviation to the mean) is a standardised dimensionless measure of dispersion. A robust counterpart is the ratio of the IQR to the median. Another alternative measure of dispersion is the median absolute deviation, defined as the median of the absolute deviations from the median. The median absolute deviation is less affected by outliers than the standard deviation as points far from the centre are given less weight (in the calculation of the standard deviation, by contrast, they are squared). The mean absolute deviation (defined as the mean of the absolute deviations from the mean) is another measure of dispersion that shares the same property. Alternatively, the mean absolute difference is computed as the arithmetic mean of the absolute value of all possible differences between the observations of the variables in the data set. When this is divided by the arithmetic mean it defines the relative mean absolute difference, which is twice the Gini coefficient.

The assessment of the degree of association between two variables is typically a measure of correlation. The basic measure here is the Pearson correlation, which measures the degree of linearity of the bivariate relationship between the two variables. The Pearson correlation is a standardised covariance (i.e. the ratio of the covariance between the two variables to the product of their variances). The Pearson correlation is bounded below by –1 (a perfect negative linear relation) and above by 1 (a perfect

positive linear relation). The Pearson correlation is not affected by changes in either the scale (e.g. multiply variables by 2) or the location (e.g. add a constant) of the variables. Two independent variables have a correlation coefficient of 0. However, a finding of zero correlation between two variables can only be interpreted as indicating independence in special cases, such as the bivariate normal distribution. The measure of association for ordinal data is the polychoric correlation (Drasgow, 1988). Using the Pearson correlation would be misleading here, due to its underlying assumption of linearity. Figure 7.5 presents Anscombe's quartet, a series of four data sets constructed by Anscombe (1973) to underline the importance of visualising the data before carrying out statistical analysis. Each data set consists of 11 pairs of points with a Pearson correlation of 0.816. The mean of the first variable (on the x-axis) is always 9, with a sample variance of 11. The mean of the second variable is approximatively 7.50, with a sample variance of between 4.122 and 4.127. Pearson correlation makes sense only in panel (a), for normally distributed data with a linear relationship. In panel (b) the relationship is clearly non-linear. Panels (c) and (d) show how the Pearson correlation is sensitive to outliers. In panel (c), without the outlier, the correlation is 1. One single outlier suffices to reduce this to 0.816. In panel (d), the correlation without the outlier is 0 but one single outlier suffices to increase this to 0.816.

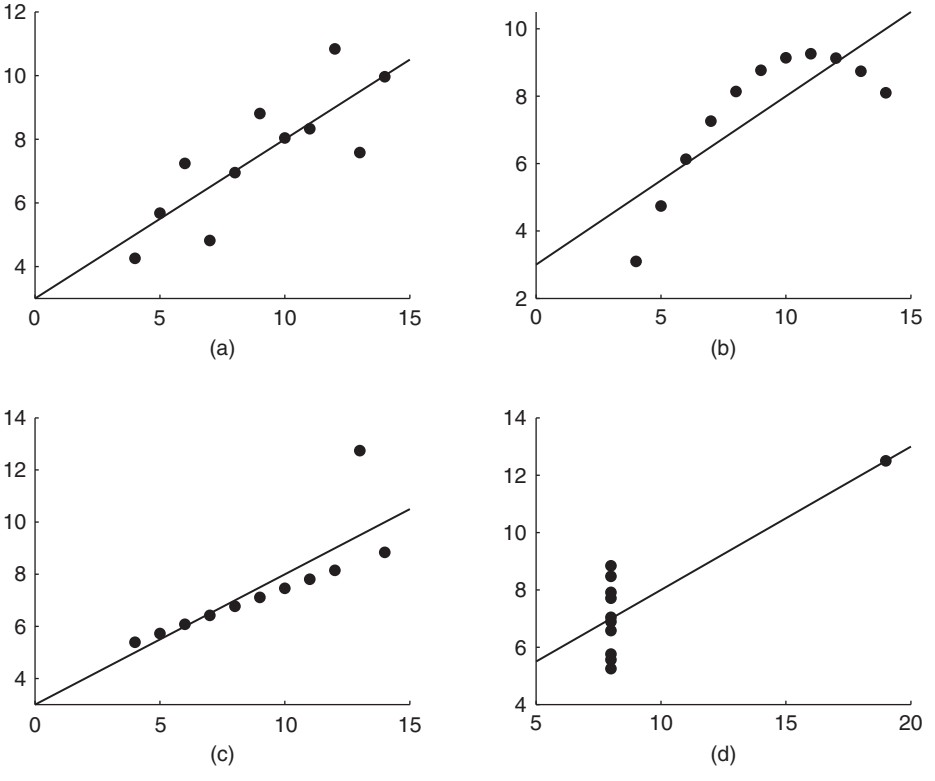


Figure 7.5 Anscombe's quartet

Table 7.3 Correlation measures and the Anscombe quartet

Measure	Panel (a)	Panel (b)	Panel (c)	Panel (d)
Pearson	0.816	0.816	0.816	0.816
Spearman	0.818	0.691	0.991	0.500
Kendall	0.636	0.564	0.964	0.426
Hoeffding's D	0.220	0.390	1	−0.150
Distance correlation	0.824	0.869	0.906	0.807

Note. Correlation measures computed on the Anscombe quartet data, displayed in Figure 7.5. Each column refers to the corresponding panel in the figure.

An alternative to the Pearson correlation coefficient is Spearman's ρ, which is the Pearson correlation of the ranks in the data. As it is based on ranks, Spearman's rho measures the tendency for two variables to rise together, without assuming that the increase is represented by a linear relationship. Spearman's rho is particularly useful when the variables are non-normally distributed. In the same vein, Kendall's τ coefficient is a non-parametric measure of rank correlation. Some alternatives to these traditional measures of association include Hoeffding's D and distance correlation. Focus 7.2 describes distance correlation as a measure of the degree of association. Hoeffding's D is robust to non-monotonic relationships. Unlike traditional measures of association, the distance correlation is 0 if and only if the random variables are statistically independent. Table 7.3 shows the values taken by these correlation measures for each data set in Anscombe's quartet.[2] All measures appear to perform better in this example than the Pearson correlation coefficient. Kendall's τ coefficient also appears to be a satisfactory alternative measure.

Experimental designs often consist in comparing individual behaviour under different treatments. When one of two possible outcomes is measured, and there is a supposed causal factor, odds ratios can be used as a simple measure of the strength of the relationship between the cause and its consequence. Focus 7.3 describes the use of odds ratio in exploratory analysis of treatment effects.

7.1.5 Methods for the Analysis of Experimental Data

There are two broad types of analytical method for experimental data. Parametric methods are based on parametrised families of probability distributions (e.g. normal, lognormal and Poisson); they apply to situations in which the assumed distribution matches the one that generated the data. Non-parametric methods, by contrast, do not assume any specific underlying distribution. The choice between the two is not always easy. Parametric methods are more powerful, in a statistical sense, thanks to the assumed knowledge about the data-generating process; while non parametric methods are more robust thanks to the lack of such assumption.

Non-parametric analysis is useful when it is known that the variable of interest is not normally distributed (or more generally does not have a known distribution). This is

[2] The Hoeffding's D value in the table is calculated with the *hoeffd* function in R. It is 30 times Hoeffding's original D, and ranges from –0.5 to 1.0 if there are no ties in the data.

Focus 7.2

Distance correlation as a measure of the degree of association

The distance correlation, denoted dCor, is the ratio of the distance covariance to the product of the distance standard deviations of the variables.

$$\mathrm{dCor}(X, Y) = \frac{\mathrm{dCov}(X, Y)}{\sqrt{\mathrm{dVar(X)}\,\mathrm{dVar(Y)}}}$$

The calculation of the distance covariance $\mathrm{dCov}(X, Y)$ and distance variances $\mathrm{dVar}(X)$ and $\mathrm{dVar}(Y)$ use the matrices of doubly centred distances (i.e. the matrices of Euclidean distances with row/column means subtracted and the grand mean added). These matrices are constructed by calculating all pairwise distances (with the Euclidean norm) in each distribution $a_{i,j} = ||X_i - X_k||$ for variable X and $b_{i,j} = ||Y_i - Y_k||$ for variable Y, $\forall i, j = 1, \ldots, n$. Doubly centred distances are obtained by subtracting the column mean (over j, denoted $\bar{a}_{i.}$) and the row mean (over i, denoted $\bar{a}_{.j}$) from each pairwise distance and adding the grand mean of the pairwise distances. The elements of the doubly centred distance matrices are denoted $A_{i,j}$ for X and $B_{i,j}$ for Y. For instance, the doubly centred distances for variable X are defined as:

$$A_{i,j} = a_{i,j} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$$

The squared sample distance covariance is defined as the arithmetic average of the products $A_{i,j}$ and $B_{i,j}$: $dCov^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{i,j} B_{i,j}$ and the distance variances are the square roots of $dCov^2(X, X)$ and $dCov^2(Y, Y)$. The distance correlation ranges from 0 to 1. It is 0 if and only if X and Y are independent. With bivariate normal distributions, it is always less than or equal to the absolute value of the Pearson correlation, and is equal to 1 when the absolute value of the Pearson correlation is one.

typically the case with ordinal variables. Parametric assumptions are also hard to justify when there are obvious outliers in the data. The assumption of normality is also harder to defend in small samples. The sample size over which violations of normality should not cause major problems is traditionally considered to be 30–40.[3]

Non-parametric methods are useful for the univariate analysis of aggregate data. They can also be used to identify differences between treatments or sessions, with fewer assumptions than under parametric methods. This does not, however, mean that non-parametric methods make no assumptions about the structure of the data. For example, non-parametric tests of the difference between two treatments are distribution-free, but still assume that the distribution is the same for the two treatments. Another assumption is that the relevant parameters, apart from the treatment, are randomly allocated between subjects. Some specific methods, like the Wilcoxon test, impose particular assumptions, such as the symmetry of the population distribution. Non-parametric methods are often simpler and are generally more robust than parametric methods, but these advantages come at a cost. Typically, multivariate or conditional analysis rapidly becomes complex with non-parametric methods. Moreover, the estimation of marginal effects is, by the nature of the approach, not possible. With respect to these arguments, parametric

[3] The central-limit theorem states that the sampling distribution for large samples tends to be normal, regardless of the shape of the data.

Focus 7.3
The exploratory analysis of treatment effects with odds ratios

Odds are defined as the ratio of the probability that a particular event occurs to the probability that it does not occur. It can be any number between 0 and infinity. Intuitively, the odds provide a normalised measure of the size of the effect of an event: they compute the number of subjects who experience the event for every subject who does not:

$$\text{odds ratio} = \frac{\text{odds of event in the treated group}}{\text{odds of event in the control group}}$$

An alternative measure is the risk ratio, which compares the frequency of the event in the treatment group to the frequency of the event in the control group.

$$\text{risk ratio} = \frac{\text{frequency of event in the treated group}}{\text{frequency of event in the control group}}$$

As an example, Cooper and Fang (2008) compare overbidding in second-price auctions when players receive no signal about their opponents to the situation in which they receive a high-quality signal about their opponent's values. The first treatment is the control and the second the signal. The following table shows the number of subjects by treatment group and bidding behaviour.

	Control	Signal
Overbidding	19	17
No overbidding	17	33
Total	36	50

The odds ratio here is the ratio of the odds of overbidding in the treatment group ($17/33 = 0.51$) to those in the control group ($19/17 = 1.12$): $0.51/1.12 = 0.56$. Overbidding is thus more likely in the control group than in the signal group: the odds of overbidding are roughly cut by a half when players receive an informative signal about their opponents. The risk ratio is the ratio of the frequency of overbidding in the experimental group ($17/50 = 0.34$) to that in the control group ($19/36 = 0.53$): $0.34/0.53 = 0.64$. The risk ratio is the factor by which the frequency of the event (i.e. the risk) is multiplied in the signal group. Receiving an informative signal about opponents reduces the risk of overbidding by around one-third.

methods are better able to deal with complexity, as they have fewer parameters, and are more accurate.

Departures from normality do not necessarily prevent the use of parametric methods. A transformation of the data can sometimes produce normality. First, sub-parts of skewed distributions can be compressed more than other parts. A typical example is the logarithmic transformation for distributions that are bounded at 0 and right-skewed. Figure 7.6.a shows an example of a logarithmic transformation.[4] For distributions that are left-skewed the same transformation can be applied to reflected distributions. Figure 7.6.b shows an example of such a transformation with a logarithm. The square

[4] A more extreme alternative is the reciprocal transformation ($1/y$), and a less extreme alternative is the square root ($\sqrt{y}$). The square root is also often used to transform count (i.e. integer) data. In this case,

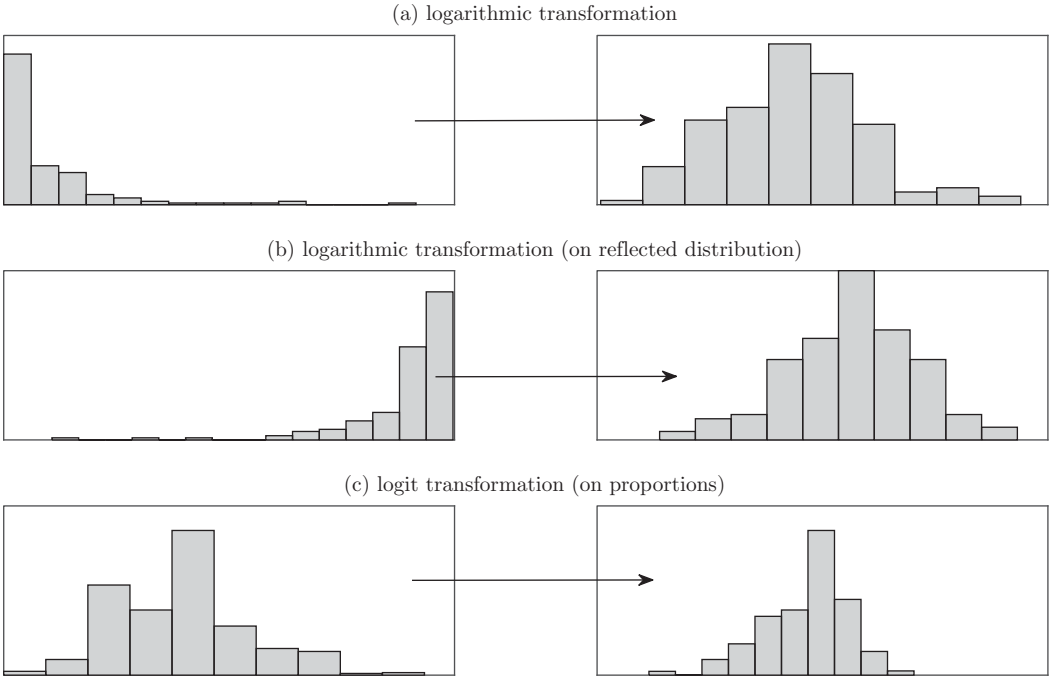


Figure 7.6 Transformation functions and normality

transformation can also be used. The logit ($\log[y/(1-y)]$) or angular ($arcsin\sqrt{y}$) transformations can be used for proportions outside the range [0.3–0.7].[5] The logit function treats very small and very large values symmetrically. It pushes out the tails and pulls in the middle of the distribution. Figure 7.6.c shows an example of a logit transformation of a distribution of proportions.

7.2 Estimation and Inference

We now turn to to the relation between the content of the data and the populations from which the sample has been drawn. To that end, denote $\mathbf{y} = (y_1, \ldots, y_n)$ the sample of observations, in which each y_i is supposed to be a draw of a random variable Y_i. In the next two sections, we assume that the elements in the samples are independent draws from the same population distribution. The cumulative distribution function of all variables Y_i is denoted G and g is the associated density. In statistical terms, the Y_is are

a common transformation is $\sqrt{y+0.4}$. A more general form of power transformation is the Box–Cox transformation:

$$\begin{cases} \frac{(y+a)^{\lambda}-1}{\lambda} \text{ if } \lambda \neq 0 \\ \log(y+a) \text{ if } \lambda = 0 \end{cases}$$

[5] The square root and log transformations are also used to reduce the variance.

assumed to be independently and identically distributed. The collection of draws in the sample informs about G because all draws arise from the same underlying distribution. The vector of random variables $(Y_1, \ldots, Y_n)$ is denoted $\mathbf{Y}$. When there are K samples, or K observations for each experimental unit i, the population distributions (if any) are denoted F_k, the samples are denoted $\mathbf{y^k} = (y_1^k, \ldots, y_{n_k}^k)$ for $k = 1, \ldots, K$ and the random variables $\mathbf{Y^k} = (Y_1^k, \ldots, Y_{n_k}^k), k = 1, \ldots, K$.

7.2.1 Estimators and Sampling Distributions

First focus on parametric analysis. We suppose that the data are generated by a parent distribution G depending on a fixed set of unknown parameters θ. Inference aims to uncover θ in the parent distribution $G(Y|\theta)$ given the observed sample $\mathbf{y}$. The central element of inference is the estimator.

Estimators

An estimator can be defined as a procedure $\hat{\theta}$, applied to the sample data, which returns a numerical value for the parameter of the parent distribution θ. The objective is to obtain values of the unknown parameters from the observations of the random variables $\mathbf{Y}$. To that end, the estimator transforms observations into parameter values; it is defined as a function $\hat{\theta}(\mathbf{Y})$ of the n random variables $Y_i, i = 1, \ldots n$.

As $\hat{\theta}$ is a function of random variables, it is itself random, with a distribution called the sampling distribution. The sampling distribution is a key element of inference, as it allows inference to be based not only on the observations $\mathbf{y}$ but also on all possible values of $\mathbf{Y}$. In frequentist estimation, applied to a given sample $\mathbf{y}$, the estimator produces a point estimate of the parameter of interest and the sampling distribution produces a confidence interval around this point estimate. The sampling distribution also allows significance tests to be carried out on the value of the underlying parameter θ.

The sampling distribution can be derived in four different ways. It can first be derived analytically and explicitly. This is the case, for example, for the sampling distribution of the sample mean estimator $\bar{Y} = \sum_i Y_i/n$ of a normal population distribution. A second possibility is to derive the large sample limit of the estimator, which is used as an approximation. This is the case for the sampling distribution of the relative frequency estimator when the product of the sample size n and the population proportion $\bar{Y} = \sum_i Y_i/n$ is greater than 10. A third possibility is the use of permutation tests. Permutation tests calculate the test statistic for all permutations (i.e. relabellings) of the original observations.[6] A last possibility is bootstrapping. While permutation tests rearrange the data according to all possible combinations, bootstrapping is based on resampling with replacement from the whole sample. If only a subset of the sample is used to perform the resampling, this technique is called the *jacknife*. In a nutshell, the bootstrap resamples from the initial sample to produce more data. Bootstrapping is usually carried out to obtain the whole sampling distribution. Jackknife, by contrast, is mainly used to calculate the variance of an estimator based on the following steps:

[6] For the permutation to be valid, *exchangeability* is required: the joint distribution of the permutations, under the null hypothesis of the test, remains unchanged. A typical case of exchangeability is *i.i.d.* data.

resample subsets of the sample of size $n-1$ without observation i, calculate the estimator $\hat{\theta}_{-i}$ for each of these, and then calculate the variance of these estimators over the n resamplings. The mean of the $\hat{\theta}_{-i}$s can also be used to estimate the bias of an estimator $\hat{\theta}$. As the removal of one observation should change the data set only a little, the jackknife is also often used to check the data set. The data set passes the test if a number of resamplings yield similar summaries of the data.

An estimator results from an objective function, defined over the true parameter, θ, and some prior information about the data-generating process. Several methods can be used to define the objective function, resulting in different estimation methods. The method of moments matches the sample moments to the corresponding distribution moments. The maximum-likelihood estimator (MLE) selects the value of θ that maximises the joint probability of a set of observations: $Pr[\mathbf{Y} = \mathbf{y}|\theta]$. If the Y_is are i.i.d., this amounts to maximising the joint product $\prod_i Pr[Y_i = y_i|\theta]$. Estimators can also be obtained by Bayesian methods, by considering the expectation of the posterior distribution.

The Quality of the Estimator

Performing satisfactory inference requires a good estimator. The quality of the estimator is closely linked to the properties of the sampling distribution. The aim is to select an estimator whose sampling distribution is as informative as possible about the true value of the parameter of interest θ. This results in three main properties defining the quality of an estimator $\hat{\theta}$:

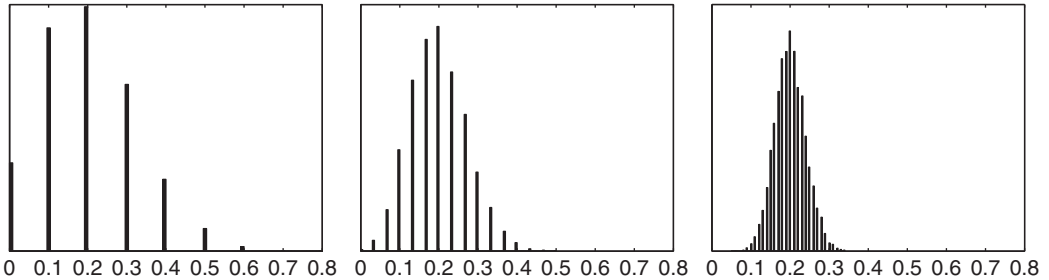
- **Consistency**. It is closer and closer to the true value θ as the sample size increases.
- **Unbiasedness**. Its expectation equals the true value, for any sample size: $\mathbb{E}(\hat{\theta}) = \theta$. This property implies that the sampling distribution of the estimator is located, on average, at the true parameter value θ.
- **Efficiency**. It has minimal variance. An unbiased estimator is better if it has greater precision, i.e. a sampling distribution with the smallest variance. The estimator with the lowest variance is called the most efficient estimator. Efficiency can be measured, for instance, as the ratio of the variance of two estimators.

The quality of an estimator can be summarised by the mean squared error (MSE), which can be decomposed in two parts: the sum of the square of the bias and the variance of the estimator:

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \underbrace{\left[\mathbb{E}(\hat{\theta}) - \theta\right]^2}_{\text{square of the bias}} + \underbrace{\mathbb{V}(\hat{\theta})}_{\text{variance of } \hat{\theta}}$$

As the expression shows, the choice of an estimator solves a trade-off between unbiasedness (the first term of the RHS) and precision (the second). In some cases, a biased estimator may be preferred if it has low variance, for example. In this case, efficiency comes at the price of bias, but the biased estimator has the lowest MSE as the gain in precision outweighs the lower unbiasedness.

Example: A natural estimator for the parameter of the binomial distribution $\mathrm{B}(n, \theta)$ *is the relative frequency estimator* $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, *with* $Y_i \in \{0, 1\}, \forall i$. *This estimator:*



(a) $n = 10$ (b) $n = 30$ (b) $n = 100$

Figure 7.7 An illustration of the central limit theorem

- *is both consistent and unbiased,* $\mathbb{E}(\hat{\theta}) = \frac{1}{n} \times n \times \mathbb{E}(Y_i) = \theta$*;*
- *has variance equal to* $\mathbb{V}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$*. The variance falls with the sample size, n: the estimates of* $\hat{\theta}$ *are more precise in larger samples. This is illustrated in Figure 7.7, which displays three samples of different sizes drawn from a binomial distribution with parameter* $\theta = 0.2$*. The variance clearly falls with the sample size. It is also worth noting that for* $n = 30$*, the distribution is approximatively normal, as suggested by the central limit theorem.*

An alternative to the relative frequency estimator is the last observation, $\hat{\theta} = Y_n$*. This is also an unbiased estimator of* θ*, but it is not consistent and has a larger variance than the relative frequency estimator.*

Last, the estimator $\hat{\theta} = \frac{1+\sum_{i=1}^{n} Y_i}{n+2}$ *is consistent and has lower variance, equal to* $\frac{\hat{\theta}(1-\hat{\theta})}{n(n+2)^2}$*, than the relative frequency estimator. The efficiency gain comes at the price of a bias equal to* $\frac{n\hat{\theta}+1}{n+2}$.

Maximum-likelihood estimators are often preferred as they are intuitive and consistent. Moreover, under regularity conditions they are consistent,[7] asymptotically efficient and asymptotically normal with mean θ and a covariance matrix equal to the inverse of the Fisher information matrix. MLEs have therefore the required properties to be a good estimator, but only asymptotically. MLEs are often inefficient and biased in finite samples. Generalisations of MLEs, such as M-estimators, have been proposed to deal with this problem.[8] Another possibility is to use Bayesian parameter estimation. Focus 7.4 shows how parameter estimation is carried out in this framework.

Examples of MLEs:

- *For multinomial probabilities, the relative frequency* $\hat{\theta}^k_{mle} = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$ *is the maximum-likelihood estimator. It is consistent and unbiased, but has a 'large' variance. If efficiency is a concern, then the MLE would not be the best estimator.*

[7] The regularity conditions impose that, for different true parameter values, the population distributions are different, that these distributions have common support for all θ, and that the true parameter is an interior point.

[8] M-estimation involves minimising a sum of functions of the data. M-estimators are based on loss functions that attach less weight to outliers.

Focus 7.4
Bayesian parameter estimation

According to the frequentist approach, the observed data are used to construct a point estimate of the parameter of interest. Bayesian statistics proceed differently: all the uncertainties (randomness, lack of knowledge) are embedded in the econometric model as probability distributions, which are combined thanks to probability calculus (typically using the Bayes theorem). The lack of knowledge about the parameters is formalised by beliefs $\mathcal{I}$ that describe the uncertainty about θ, before anything is observed. They capture the subjective information available with a prior distribution, $f(\theta|\mathcal{I})$, expressing what the researcher knows about the parameter before observing any data. The prior distribution is usually based on judgement from experts or on technical convenience for the calculation. Randomness is represented by the likelihood $Pr[\mathbf{y}|\theta, \mathcal{I}]$. In the Bayesian framework, parameter estimation corresponds to the calculation of the probability distribution over θ, given the observed data $\mathbf{y}$ and beliefs $\mathcal{I}$. The posterior distribution $Pr[\theta|\mathbf{y}, \mathcal{I}]$ combines the prior with the information obtained from the data according to Bayes's rule:

$$Pr[\theta|\mathbf{y}, \mathcal{I}] = \frac{Pr[\mathbf{y}|\theta, \mathcal{I}] \times f(\theta|\mathcal{I})}{Pr[y|\mathcal{I}]} = \frac{\text{likelihood for } \theta \times \text{prior over } \theta}{\text{likelihood marginalised over } \theta}$$

The numerator is the product of the standard likelihood of the parameter for the data $Pr[\mathbf{y}|\theta, \mathcal{I}]$ and the prior distribution over θ, given the beliefs. The denominator is the marginal likelihood, $\int_\theta Pr[\mathbf{y}|\theta, \mathcal{I}] \times f(\theta|\mathcal{I}) d\theta$, which can usually be ignored in calculations with normalisation. The posterior distribution allows us to make inferences about the model parameters:

- Location measures – the mode, the median and the mean of the posterior distribution given the point estimates.
- Credible intervals – the range of values that has the posterior probability $(1 - \alpha)$ of containing the parameter.
- The posterior probability for some hypothesis to be true, allowing hypothesis testing (e.g. for $H_0 : \theta \leq \theta_0$, we have $Pr[\theta \leq \theta_0|\mathbf{y}] = \int_{-\infty}^{\theta_0} Pr[\theta|\mathbf{y}, \mathcal{I}] d\theta$).

The table below provides a comparison between the frequentist and Bayesian methods.

	Frequentist	Bayesian
Parameter	Unknown constant	Random variable
Point estimation	Value of an estimator	Posterior summary
Interval estimation	Confidence interval	Credible interval
Hypothesis testing	Check if estimate is probable given the sampling distribution under H_0	Posterior probability of H_0

A simple example is the estimation of a binomial proportion. With a uniform prior, the posterior distribution is a $Beta(a', b')$ distribution with $a' = \sum y_i + 1$ and $b' = n - \sum y_i + 1$. With a prior $Beta(a, b)$, the posterior is a Beta distribution with parameters $a' = a + \sum_i y_i$ and $b' = b + n - \sum_i y_i$. Bayesian updating implies updating parameter a with the number of successes, and parameter b with the number of failures. For the estimation of a normal

posterior distribution of mean $\bar{Y}$ and variance σ^2/n. With a normal prior $\mathcal{N}(m, s^2)$, the posterior is normally distributed

$$\mathcal{N}\left(\frac{\frac{\sigma^2}{n}m + s^2\bar{Y}}{\frac{\sigma^2}{n} + s^2}, \frac{\sigma^2 s^2}{ns^2 + \sigma^2}\right)$$

The mean is a weighted average of the prior mean m and the sample mean $\bar{y}$. Gelman et al. (2014) provide an extensive overview of Bayesian parameter estimation.

- *For the mean of a normal distribution, the sample mean $\bar{Y} = \hat{\theta}_{mle} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the maximum-likelihood estimator. It is consistent, unbiased and efficient.*
- *For the variance of a normal distribution, the sample variance divided by n is the maximum-likelihood estimator $\hat{\theta}_{mle} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. It is consistent but biased (the MLE is said to 'over-fit' the data). To eliminate this bias, the likelihood is adjusted by a factor of $n/(n-1)$. This adjustment leads to the sample variance:*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

which is different from the maximum-likelihood estimator of the variance.

7.2.2 Confidence Intervals

Point estimates do not inform about the confidence the researcher can have in the sample value of the estimator (the point estimate). The confidence in a point estimate is by nature limited, as the sample contains less information than the population. In frequentist estimation, the population parameter is deterministic, fixed, but unfortunately unknown. The only resource available to the researcher is the estimator, which is a random variable, and the point estimate is nothing more than a draw of this random variable. As such, point estimates taken from samples always involve some uncertainty, arising from the sampling distribution. A confidence interval quantifies the uncertainty surrounding an estimate. The logic is to make confidence statements by calculating the interval that would contain the true value were the sampling to be repeated $100 \times \alpha$ times. Formally, a confidence interval is a random interval $[b_L; b_U]$ such that:

$$Pr[b_L \leq \theta \leq b_U] = 1 - \alpha$$

Note that the confidence interval tells us nothing about the location of the true parameter within the interval. Figure 7.8 illustrates the logic of confidence intervals with symmetric intervals.[9] Multiple $100 \times (1-\alpha)\%$ confidence intervals are represented from different samples, each from the same population with an unknown true parameter θ. In Figure 7.8, about 19 of the 20 confidence intervals contain the true parameter value θ.

Confidence Intervals for the Mean

Suppose that the population distribution is (approximatively) normally distributed with mean θ, the unknown parameter to be estimated, and a known standard deviation of σ.

[9] The confidence intervals $[b_L; b_U]$ are not necessarily symmetric.

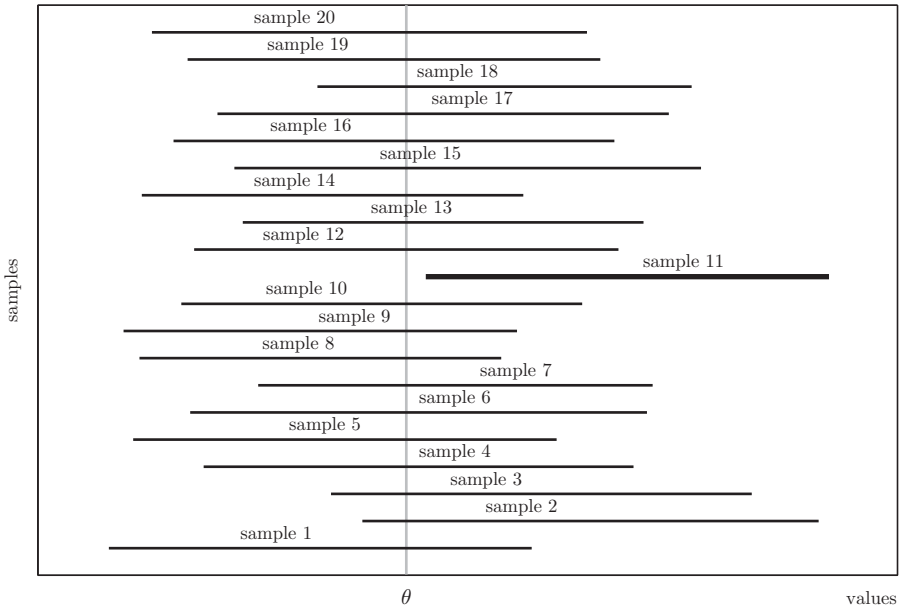


Figure 7.8 Confidence intervals on samples from a population with parameter θ

As each Y_i is supposed to be *i.i.d.* drawn from $\mathcal{N}(\theta, \sigma^2)$, the sample mean $\bar{Y}$ is distributed $\mathcal{N}(\theta, \sigma^2/n)$ and:

$$Pr\left[-z_{\frac{\alpha}{2}} \leq \frac{\hat{\theta} - \theta}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

where $z_{\frac{\alpha}{2}}$ denotes the critical value of the normal distribution such that $\Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$. Figure 7.9 shows where the critical values $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ are located for a standard normal. $z_{\frac{\alpha}{2}}$ is also called the z-score. The confidence interval $[b_L; b_U]$ for the unknown parameter θ is therefore:

$$\left[\bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

The z-score measures to what extent, in terms of standard deviations, the bounds of the confidence interval are above and below the mean. The z-score depends on the desired level of confidence: for $1-\alpha = 95\%, z_{\frac{\alpha}{2}} = 1.96$, for $1-\alpha = 90\%, z_{\frac{\alpha}{2}} = 1.645$, for $1 - \alpha = 99\%, z_{\frac{\alpha}{2}} = 2.58$. As the z-score rises with confidence, higher levels of confidence result in larger confidence intervals. On the contrary, larger sample sizes yield narrower confidence intervals. Very large samples greatly reduce the uncertainty surrounding the point estimate.

So far, we have described the bounds of the confidence intervals as random variables. For a given sample $\mathbf{y}$, the point estimate is $\bar{y}$ and the confidence interval is $\bar{y} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. If sampling is carried out without replacement from a population of size N, with N less

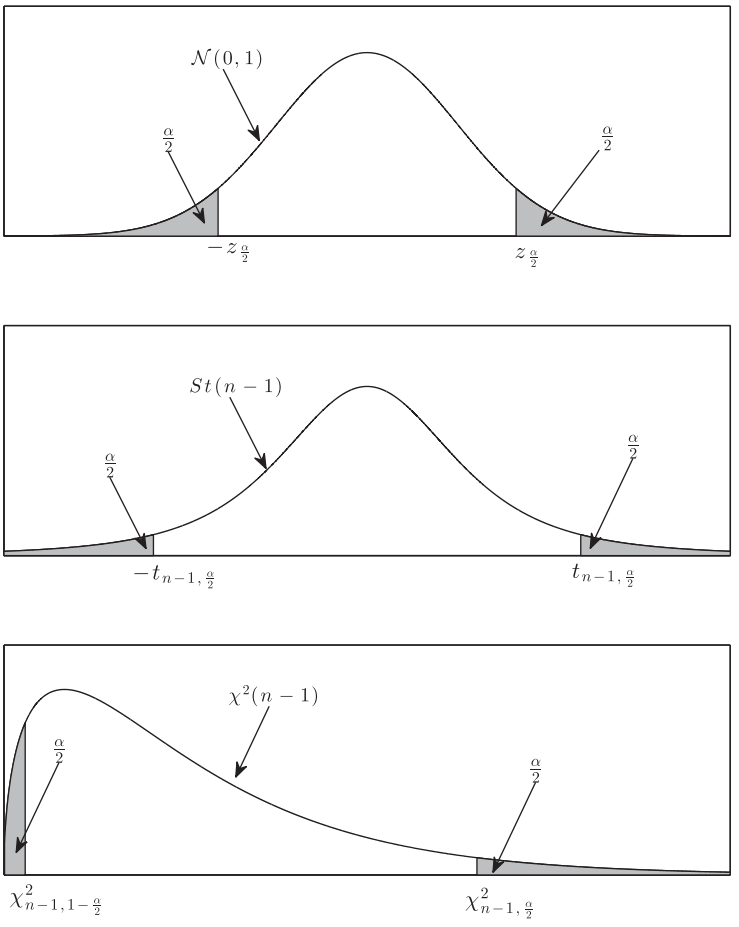


Figure 7.9 Critical values for common distributions: normal, *t* and χ^2

than 10 to 20 times n, a finite population correction needs to be applied to the standard error of the mean $\frac{\sigma}{\sqrt{n}}$, and the confidence interval becomes:

$$\left[\bar{y} - z_{\frac{\alpha}{2}} \underbrace{\sqrt{\frac{N-n}{N-1}}}_{\textbf{correction}} \frac{\sigma}{\sqrt{n}}; \bar{y} + z_{\frac{\alpha}{2}} \underbrace{\sqrt{\frac{N-n}{N-1}}}_{\textbf{correction}} \frac{\sigma}{\sqrt{n}}\right]$$

When the population standard deviation σ is unknown, the standard deviation of the sampling distribution cannot be calculated, and the sample standard deviation has to be used. The standard deviation of the sample is the square root of the sample variance s^2, defined as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

In this case, for a large sample size n, the sample mean $\bar{Y}$ is approximatively distributed $\mathcal{N}(\theta, s^2/n)$ and the confidence interval $[b_L; b_U]$ for the unknown parameter θ is therefore:

$$\left[\bar{y} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}; \bar{y} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

If the sample size is small, a normal approximation can no longer be used for the sampling distribution. A t-distribution with $n-1$ degrees of freedom is instead used to approximate the sampling distribution of $\frac{\hat{\theta}-\theta}{\frac{s}{\sqrt{n}}}$, and the confidence interval $[b_L; b_U]$ for the unknown parameter θ is therefore:

$$\left[\bar{y} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}; \bar{y} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

where $t_{\frac{\alpha}{2}}$ is the t-score with $n-1$ degrees of freedom and a decumulative probability of $\frac{\alpha}{2}$. The t-score is the critical value such that a t-distribution with $n-1$ degrees of freedom has a cumulative probability of $1-\frac{\alpha}{2}$. Figure 7.9 shows the critical values $-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$ for a t-distribution with $n-1$ degrees of freedom. Figure 7.9 also shows that the t-distribution has heavier tails than the standard normal, and that the $t_{\frac{\alpha}{2}}$ are higher than the corresponding z-scores.[10] If sampling is carried out without replacement the finite population correction needs to be applied to the sample standard error $\frac{s}{\sqrt{n}}$.

Rather than calculating the confidence interval for an unknown parameter for a given sample and a given level of confidence, an alternative is to determine the sample size that is necessary to obtain a given width for the confidence interval. Focus 7.5 shows how sample size and width at a given level of confidence relate. The method used to construct confidence intervals can also be used to predict values. Focus 7.6 shows how it is possible to predict a single value of a variable given a sample.

Confidence Intervals for the Median

When the population distribution is not normal, calculating a confidence interval for the median is a good alternative to calculating a confidence interval for the mean. Here the n observations are ordered by size, with the ordered sample values being written as $y_{(1)} \leq y_{(2)} \leq .... \leq y_{(n)}$. An equal-tailed confidence interval for the median is $[Y_{(h)}; Y_{(h')}]$, where h is the largest integer, and h' the smallest, such that:

$$\sum_{i=0}^{h-1} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2} \text{ and } \sum_{i=h'}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \frac{\alpha}{2}$$

For $n > 50$, a normal approximation of the binomial distribution can be used, and the confidence interval for the median is $[Y_{(h)}; Y_{(h')}]$, where h and h' are approximated by:

$$h \approx \frac{n+1-z_{\frac{\alpha}{2}}\sqrt{n}}{2} \text{ and } h' \approx \frac{n+1+z_{\frac{\alpha}{2}}\sqrt{n}}{2}$$

[10] As n tends to infinity, the t-score converges on the z-score.

Focus 7.5
Sample size and confidence intervals

Using confidence intervals, it is possible to compute the sample size necessary to obtain a given width at a given level of confidence. Suppose that we want to estimate the mean θ of a normal population with known variance σ. In this case, the confidence interval for a given sample **y** is

$$\left[\bar{y} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; \bar{y} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

Its width, defined as the range between the lower and the upper bound, is $width = 2z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$. The width can be interpreted as the error margin for the unknown parameter. The sample size n necessary to obtain a given value of *width* around the sample mean $\bar{y}$ at a confidence level of $100(1-\alpha)\%$ is therefore:

$$n = \left(2z_{\frac{\alpha}{2}} \frac{\sigma}{width}\right)^2$$

Analogous reasoning produces the sample size n necessary to obtain a given value of *width* around a sample frequency $\bar{y}$ at a confidence level of $100(1-\alpha)\%$:

$$n = \left(2z_{\frac{\alpha}{2}} \frac{\bar{y}(1-\bar{y})}{width}\right)^2$$

The number of independent observations required to assess a difference Δ between the means of two normally distributed populations with identical variance σ is

$$n = 2(z_{\frac{\alpha}{2}} + z_{beta})^2 \left(\frac{\sigma}{\Delta}\right)^2$$

with β being the Type-II error.

Note that the confidence interval for the median can be generalised to any quantile. An equal-tailed confidence interval for the p^{th} quantile is $[Y_{(h)}; Y_{(h')}]$, where h is the largest integer and h' is the smallest integer, such that:

$$\sum_{i=0}^{h-1} \binom{n}{i} (p)^i (1-p)^{n-i} \leq \frac{\alpha}{2} \text{ and } \sum_{i=0}^{h'-1} \binom{n}{i} (p)^i (1-p)^{n-i} \geq 1 - \frac{\alpha}{2}$$

Confidence Intervals for a Proportion

Suppose that the population is distributed according to a Bernoulli distribution with parameter θ, the unknown proportion to be estimated. The estimator $\hat{\theta}$ of θ is the relative frequency $\bar{Y} = \sum_i Y_i/n$. As each Y_i is supposed to be *i.i.d.* drawn from the Bernoulli distribution, the sum of the Y_i over $i = 1, \ldots, n$ is binomially distributed $B(n, \theta)$. If the sample size is not too small and the number of successes and failures in the sample are both large, $\sum_i y_i > 5$ and $\sum_i (1 - y_i) > 5$, a normal approximation can be used to represent the sampling distribution. In this case, the relative frequency is approximatively distributed $\mathcal{N}[\theta, \theta(1-\theta)/n]$. The confidence interval is

Focus 7.6
Prediction intervals for a single observation

The method used to construct confidence intervals can be used to predict a single value of a variable Y_{n+1} given a sample $(Y_1, \ldots, Y_n)$. For example, if the population is normally distributed with unknown mean θ and known variance σ, the sampling distribution of Y_{n+1} given $\mathbf{Y}$ is normally distributed $\mathcal{N}(\bar{Y}, \sigma\sqrt{1+\frac{1}{n}})$. A prediction interval with level $100(1-\alpha)\%$ is such that

$$Pr\left[-z_{\frac{\alpha}{2}} \leq \frac{Y_{n+1}-\bar{Y}}{\sigma\sqrt{1+\frac{1}{n}}} \leq z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

And the prediction interval is

$$\left[\bar{y} - z_{\frac{\alpha}{2}}\sigma\sqrt{1+\frac{1}{n}}; \bar{y} + z_{\frac{\alpha}{2}}\sigma\sqrt{1+\frac{1}{n}}\right]$$

$$\left[\bar{y} - \frac{1}{2n} - z_{\frac{\alpha}{2}}\frac{\bar{y}(1-\bar{y})}{\sqrt{n}}; \bar{y} + \frac{1}{2n} + z_{\frac{\alpha}{2}}\frac{\bar{y}(1-\bar{y}}{\sqrt{n}}\right]$$

As the estimator is based on discrete values and the approximation is based on a continuous distribution, a continuity correction $\frac{1}{2n}$ is needed. In the case of sampling without replacement, the finite population correction $\sqrt{\frac{N-n}{N-1}}$ needs to be applied to the standard error $\frac{\bar{y}(1-\bar{y})}{\sqrt{n}}$.

The normal approximation used to calculate the confidence interval can be problematic if the sample size is too small. The conditions imposed on the number of successes and failures are not problematic when the population proportion is close to one-half. In that case, the sample size required to produce at least five successes and at least five failures is close to 10. But if the population proportion is more extreme (close to 0 or 1), the sample size required to meet the condition of at least five successes and five failures is much larger. For example, if the population proportion is 0.01, a minimum sample size of 500 is needed. Moreover, the normal approximation assumes a symmetric margin of error that may be problematic when the estimator takes a value of 0 or 1. Instead of using a normal approximation to the sampling distribution, the correspondence between the binomial distribution and the Fisher distribution can be used. Here the confidence interval for the parameter θ is

$$\left[\frac{n\bar{y}}{n\bar{y} + (n - n\bar{y} + 1)\,\mathrm{F}_l}; \frac{(n\bar{y}+1)\,\mathrm{F}_u}{n(1-\bar{y}) + (n\bar{y}+1)\,\mathrm{F}_u}\right]$$

With $\mathrm{F}_l = F_{2(n-n\bar{y}+1),2n\bar{y},\frac{\alpha}{2}}$ and $\mathrm{F}_u = F_{2(n\bar{y}+1),2n(1-\bar{y}),\frac{\alpha}{2}}$ being the critical values of the Fisher distribution. This alternative confidence interval is very useful for extreme values. For example, the one-tailed confidence interval for complete failure is

$$\left[0; \frac{F_{2,2n,\frac{\alpha}{2}}}{n + F_{2,2n,\frac{\alpha}{2}}}\right]$$

and the one-sided confidence interval for complete success is

$$\left[\frac{n}{n + F_{2,2n,\frac{\alpha}{2}}}; 1\right]$$

Confidence Intervals for Variance and Standard Deviation

Suppose the population is (approximatively) normally distributed with mean μ and an unknown standard deviation θ. The estimator of the variance θ^2 is the sample variance S^2. As each Y_i is supposed to be *i.i.d.* drawn from $\mathcal{N}(\mu, \theta^2)$, the random variable $\frac{(n-1)S^2}{\theta^2}$ is distributed chi-squared with $n-1$ degrees of freedom. The resulting interval with confidence level $100(1-\alpha)\%$ is

$$Pr\left[\chi^2_{n-1,1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\theta^2} < \chi^2_{n-1,\frac{\alpha}{2}}\right] = 1 - \alpha$$

where $\chi^2_{n-1,1-\frac{\alpha}{2}}$ and $\chi^2_{n-1,\frac{\alpha}{2}}$ are the critical values of the chi-squared distribution with $n-1$ degrees of freedom. Figure 7.9 depicts these critical values for a chi-squared distribution with $n-1$ degrees of freedom. The confidence interval $[b_L; b_U]$ for the unknown parameter θ is therefore:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}}\right]$$

For small sample sizes and distributions with extreme values or missing data, the mean absolute deviation (MD) is a better estimate of dispersion than the standard deviation (Tukey, 1960). The mean absolute deviation is

$$MD = \frac{\sum_i |Y_i - \bar{Y}|}{n}$$

and can be used to construct confidence intervals for the mean (Sachs, 2012). The median deviation, defined as the median of the absolute deviation from the empirical mean, is another robust estimate of dispersion.

Confidence Intervals for a Difference Between Two Populations

Now suppose that two samples of observations are available: $\mathbf{y^1} = (y^1_1, \ldots, y^1_{n_1})$ and $\mathbf{y^2} = (y^2_1, \ldots, y^2_{n_2})$, which are drawn independently. Further assume that the elements of each sample are independent draws from the same population distributions: G_1 for the Y^1_i and G_2 for the Y^2_i. The confidence intervals for the difference between two independent samples is a first way of evaluating treatment effects, when treatments are applied to different experimental units.[11]

The estimator for the difference between the means of two independent normally distributed populations is $\hat{\theta} = \bar{Y}^1 - \bar{Y}^2$. If the variances of the two populations are assumed to be equal, the confidence interval for the difference between the means is

$$(\bar{y}^1 - \bar{y}^2) \pm t_{\frac{\alpha}{2}} s\sqrt{1/n_1 + 1/n_2}$$

[11] If the draws are not entirely random, the sample standard deviation will be spuriously lower, so the significance of the difference between the means will be larger.

where the estimator of the common variance is $s^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}$ and t is the t-score with $n_1 + n_2 - 2$ degrees of freedom. If the variances of the two populations are not assumed to be equal, the confidence interval for the difference between the means is

$$(\bar{y}^1 - \bar{y}^2) \pm t_{\frac{\alpha}{2}} \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

where t is the t-score with df degrees of freedom and:

$$df = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Bigg/ \left[\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right]$$

The estimator of the difference between the proportions of two independently distributed populations is the difference between the two sample frequencies $\bar{y}^1$ and $\bar{y}^2$. If $n_1 \geq 50, n_2 \geq 50$ and the expected number of successes and failures in both samples is greater than five, the sampling distribution of the estimator can be approximated by a normal distribution. In this case, the confidence interval for the difference between the proportions is

$$(\bar{y}^1 - \bar{y}^2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{y}^1(1-\bar{y}^1)}{n_1} + \frac{\bar{y}^2(1-\bar{y}^2)}{n_2}}$$

The estimator for the ratios between the variances of two independently distributed populations is $\frac{S_1^2}{S_2^2}$, where S_1^2 corresponds to the larger sampling variance. The sampling distribution of the estimator has a Fisher distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. The confidence interval $[b_L; b_U]$ for the ratio between the variances is therefore:

$$\left[\frac{s_1^2}{s_2^2}\frac{1}{F_{(n_1-1,n_2-1),\frac{\alpha}{2}}}; \frac{s_1^2}{s_2^2}F_{(n_2-1,n_1-1);\frac{\alpha}{2}}\right]$$

Confidence Intervals for Paired Data

Paired data offer a simple way of identifying treatment effects. Paired data allow a more accurate comparison between treatments as the dispersion between different experimental units is reduced, by removing the inter-individual heterogeneity. The only dispersion that remains is within-subject. On the other hand, compared to independent samples, paired data reduce the number of degrees of freedom by a factor of two. The reduction in the degrees of freedom increases the confidence intervals. As such, we are faced with a trade-off between precision and accuracy when comparing paired and independent samples. Note that pairing is a design strategy, which must be chosen before the experiment, and not an analytical strategy.

Let $\mathbf{y^1} = (y_1^1, \ldots, y_n^1)$ and $\mathbf{y^2} = (y_1^2, \ldots, y_n^2)$ be two samples of paired data. Each pair y_i^1, y_i^2 is generated by the same observation unit i. We assume that the differences $\Delta_i = Y_i^1 - Y_i^2$ are drawn at random from a normally distributed population.[12] To

[12] When comparing paired samples, it is therefore not necessary to assume that both distributions are approximatively normal, but only that their difference is approximatively normal.

obtain a confidence interval for the mean differences of the treatment effect, consider the estimator defined by the mean difference between treatments $\bar{\Delta} = \frac{1}{n}\sum_i \Delta_i$. Here the confidence interval for the mean difference in paired observations is

$$\left[\bar{\Delta} - t_{\frac{\alpha}{2}} s_\Delta; \bar{\Delta} + t_{\frac{\alpha}{2}} s_\Delta\right]$$

where $s_\Delta^2 = \frac{\sum_i (\Delta_i - \bar{\Delta})^2}{(n-1)}$ is the sample standard deviation and t is the t-score with $n-1$ degrees of freedom.

7.3 Testing Procedures

This section will review non-parametric and parametric testing procedures. We start by setting out the general principles of hypothesis testing, and then turn to the main hypothesis tests for each type of measurement. We present a number of examples of hypothesis testing in experimental economics.

7.3.1 Hypothesis Testing: General Principles

The general purpose of hypothesis testing is to make decisions based on inferences about populations from given samples. To ease exposition, we focus on parametric analysis. The particularity of hypothesis testing, as compared to measurement via point estimates and confidence intervals, is to make a judgement about some population characteristic. To make a judgement about a parameter value, hypothesis testing sets this judgement up as a testing problem between two competing, mutually exclusive, hypotheses regarding the true parameter value. A test is a rule enabling a decision to be made about whether the hypothesis under test should be rejected, based on the observed value of some function of the sample variables. Focus 7.7 summarises the approach to hypothesis testing in a five-step procedure. It should be emphasised that hypothesis testing is not a substitute for confidence intervals. Hypothesis testing makes sense if a well-defined research hypothesis exists and needs to be tested. Otherwise, it is better to use confidence intervals to summarise the information about the population that the sample contains.

The decision rule is to reject the hypothesis under investigation when the result from the analysis of the sample is sufficiently unlikely under this hypothesis. Carrying out hypothesis tests is a two-stage procedure. It is first assumed that chance alone determines the outcome under each competing hypothesis. Second, the test produces the chances of observing the various outcomes, given these assumptions. A significance level is used to determine what is probable and what is unlikely, and at which level the hypothesis may be rejected.

Carrying out a test implies making a **decision** regarding the value of θ based on the sample data $(y_1, \ldots, y_n)$ between two competing hypotheses. The first hypothesis, H_0, is called the null hypothesis. This hypothesis often corresponds to the default or 'no-effect' situation. The second hypothesis, which is usually the research hypothesis under investigation, is denoted H_1 (or H_a) and called the alternative hypothesis. There are three types of alternative hypothesis H_1, as follows:

Focus 7.7
A five-step approach to hypothesis testing

Hypothesis testing is usually carried out using the following five-step approach:

1. Set up the hypothesis and select the level of significance α.
2. Select the appropriate test statistic (and underlying estimator) $\mathbf{T}(\mathbf{Y})$, according to:
   - the level of measurement of the data (interval, ratio, ordinal or nominal);
   - the characteristics of the distributions (sample sizes, normal approximations, equal variances, etc.);
   - the design of the experiment (repeated measures, matched participants, independent populations, etc.).
3. Set up a decision rule, i.e. a statement about circumstances under which the null hypothesis is rejected. There are different types of decision rule. For some tests, the decision rule will be to reject the null hypothesis if the value of the test statistic is large; for other tests, the decision rule will be to reject the null hypothesis if the value of the test statistic is low, or outside some given bounds. If p-values are used, the null hypothesis will be rejected if the p-value is too low.
4. Calculate the sample value of the test statistic using the sample values $\mathbf{y}$.
5. Make a decision using the decision rule: to reject (decision d_1) or not to reject (decision d_0) the null hypothesis.

- The true parameter θ is lower than a certain value denoted θ_0: $\theta < \theta_0$. When tested against the null hypothesis $H_0 : \theta = \theta_0$, this test is called a left-tailed test.
- The true parameter θ is greater than a certain value denoted θ_0: $\theta > \theta_0$. When tested against the null hypothesis $H_0 : \theta = \theta_0$, this test is called a right-tailed test.
- The true parameter θ is different from a certain value denoted θ_0: $\theta \neq \theta_0$. When tested against the null hypothesis $H_0 : \theta = \theta_0$, this test is called a two-tailed test.

The first two alternatives are single-sided tests, and the third a two-sided test. In all three alternatives, the hypothesis corresponds to partitions of the parameter space – i.e. the space of possible parameter values. If the parameter space for the parameter θ is Θ, the null hypothesis and the alternative hypothesis form a partition of Θ between Θ_0 and its complement Θ_1:

$$H_0 : \theta \in \Theta_0 \subset \Theta \text{ vs. } H_1 : \theta \in \Theta_1 \subset \Theta$$

If the null hypothesis H_0 implies complete knowledge of the population distribution, the hypothesis is said to be simple. Otherwise, the hypothesis is said to be composite.

Examples:

- *With $Y_i \sim \mathcal{N}(\theta_0, 1)$, the null hypothesis $H_0 : \theta = \theta_0$ is a simple hypothesis, whereas $H_0 : \theta \leq \theta_0$ is a composite hypothesis.*
- *With $Y_i \sim \mathcal{N}(\theta, \sigma)$ and σ unknown, the null hypothesis $H_0 : \theta = \theta_0$ is a composite hypothesis because the hypothesis does not imply complete knowledge of the population distribution.*

Table 7.4 True data-generating process and decisions

DGP	d_0	d_1
H_0	No error	False positive. Type-I error
H_1	False negative. Type-II error	No error

The logic of hypothesis testing is to decide whether to reject the null hypothesis (formulated in terms of the population parameter θ) based on the sample data $\mathbf{y} = (y_1, \ldots, y_n)$. If the sample data are consistent with the probability model specified by the null hypothesis H_0, then the decision is to not reject that hypothesis. This decision is denoted d_0. If the sample data are inconsistent with the null hypothesis H_0, then the decision is to reject that hypothesis. This decision is denoted d_1. In the latter case, the alternative hypothesis H_1 regarding the data-generating process is most likely to be true. The decision d_0 or d_1 is taken on the basis of a test statistic $\mathbf{T}(\mathbf{Y})$ and a significance level. Table 7.4 shows in which cases ($d_0 - H_0$ and $d_1 - H_1$) the decisions are made without error.

In practice, the test statistic $\mathbf{T}(\mathbf{Y})$ is a function of the estimator of the parameter of interest. For example, to test the value of the mean θ of a normally distributed population with variance σ^2, the estimator is $\bar{Y}$ and the test statistic is the normalised value of the estimator $\frac{\bar{Y}-\theta}{\sigma/\sqrt{n}}$. A test statistic $\mathbf{T}(\mathbf{Y})$ is said to be feasible when its sampling distribution under the null hypothesis $\mathcal{L}^0$ is known and when its value can be calculated from a given set of sample data $\mathbf{y}$. $\mathcal{L}^0$ is sometimes referred to as the null distribution of $\mathbf{T}(\mathbf{Y})$. Under the alternative hypothesis, H_1 describes the true DGP and the distribution of $\mathbf{T}(\mathbf{Y})$ is denoted $\mathcal{L}^1$. The significance level is defined in terms of a probability threshold (denoted α), such that a particular estimate is significant if the probability of obtaining that estimate under the null distribution is less than α.

Example: Consider a sample $\mathbf{Y}$ *such that* $Y_i \sim \mathcal{N}(\theta, \sigma^2)$*, and let* $\mathbf{T}(\mathbf{Y}) = \bar{Y}$ *be an estimator of* θ*. We know that:*

$$\frac{\mathbf{T}(\mathbf{Y}) - \theta}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

We want to test $H_0 : \theta = 2$ *against some alternative (one-sided or two-sided) hypothesis. The distribution of* $\mathbf{T}(\mathbf{Y})$ *is* $\mathcal{L}^0 \sim \mathcal{N}(2, \sigma/\sqrt{n})$ *if and only if* H_0 *is the true DGP behind observed data.*

The decision rule for the test can be seen as a partition of the set of all values of the test statistic $\mathbf{T}(\mathbf{Y})$. The subset of all values of the test statistic that led to the rejection of the null hypothesis is called the 'critical' or 'rejection' region R. The critical values of a test statistic $\mathbf{T}(\mathbf{Y})$ are the bounds of R. The bounds are set in such a way that the difference between $\mathbf{T}(\mathbf{Y})$ and H_0 is large enough for the probability that it occurred by chance to be lower than a given significance level α. The critical region is thus made of all values of $\mathbf{T}(\mathbf{Y})$ that are unlikely to be drawn from the population distribution set by H_0.

Figure 7.10 shows the three types of rejection region used in one-dimensional hypothesis testing. A bilateral rejection region prescribes rejection if the sample value of the

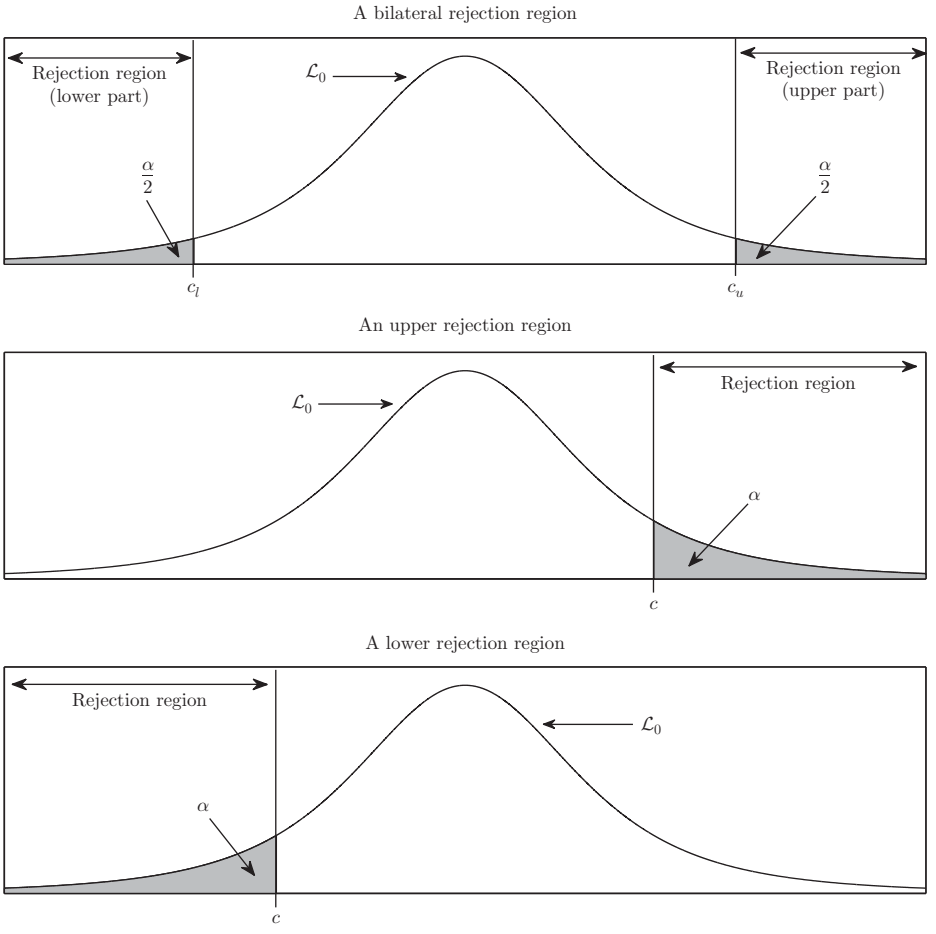


Figure 7.10 Critical values and rejection regions
Note. $\mathcal{L}_0$ denotes the sampling distribution of the test statistic T under the null hypothesis H_0.

test statistic is less than a lower threshold c_l or greater than an upper threshold c_u. An upper (lower) rejection region prescribes rejection if the sample value of the test statistic is less (greater) than a lower (upper) threshold c. Figure 7.10 shows that the rejection region R depends on three key elements:

- the alternative hypothesis H_1, one-sided or two-sided;
- the sampling distribution under the null hypothesis $\mathcal{L}_0$;
- the significance level, α.

Figure 7.11 shows graphically how hypothesis testing works. In the upper part of the figure, deductive reasoning is used by the researcher to determine the rejection region R based on the significance level α and the sampling distribution $\mathcal{L}_0$. In Figure 7.11, the hypothesis test prescribes rejection of H_0 for $\mathbf{T}(\mathbf{Y}|H_0) > c$ and the critical value c is such that $Pr[\mathbf{T}(\mathbf{Y}) > c|\mathcal{L}^0] \leq \alpha$. The decision rule states that if the sample value of

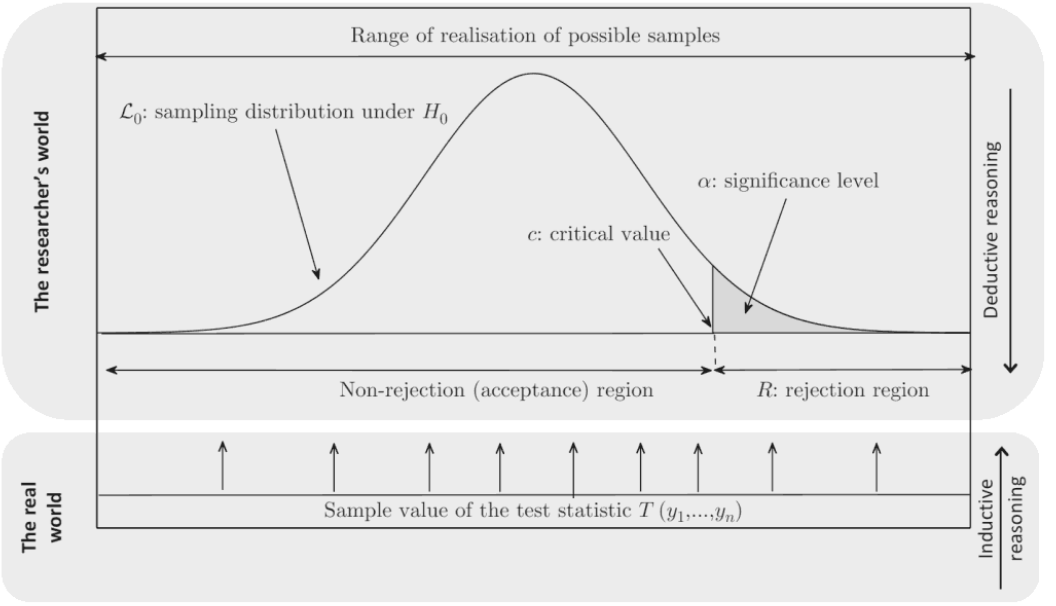


Figure 7.11 Hypothesis testing

the test statistic $\mathbf{T}(\mathbf{Y}) = \hat{\theta}$ is larger than c, then decision d_1 should be taken. Once the rejection region is determined, the researcher uses inductive reasoning on the basis of the computed sample value of the test statistic. If this value falls into the rejection region, decision d_1 is taken; if the value falls outside the rejection region, decision d_0 is taken.

As the estimator is itself a random variable, it is impossible to know for sure whether the decision (d_1 or d_0) is correct regarding the true DGP. The decision is risky to the extent that one can always draw extreme values from a distribution. In other words, the test can lead to the wrong decision. An extreme draw in the H_0 distribution leads to decide d_1. This is called a type I error, and corresponds to a false positive. False negatives arise if, by contrast, a draw in a distribution different from H_0 lies outside the critical region and leads to decide d_0. This is a type II error. Table 7.4 shows the accuracy of decisions based on the nature of the true DGP. When the decision is in line with the true DGP, the statistical decision based on the test is accurate.

Formally, the probability of a type I error is

$$\alpha(\theta) = Pr[d_1|H_0] = Pr[\mathbf{T}(\mathbf{Y}) \in R|\theta \in \Theta_0]$$

For a given choice of the level $\alpha(\theta)$, it is possible to determine the critical region R, and therefore the critical values of the test statistics, that lead to the rejection of the null hypothesis. For a simple test, α is the 'size' of the test – the largest probability of rejecting H_0 when H_0 is true. In most applications, one sample is used to carry out one test. With experimental data, however, it is common practice to use the same sample to carry out several tests. In this case, the size of the test should be modified to take

Focus 7.8
Multiple test procedures

Consider a second-price auction experiment eliciting 10 bids per subject, on which a series of 10 different tests for overbidding behaviour are performed. A level of significance $\alpha = 0.05$ implies that, just by chance, the probability that one out of the 10 tests will lead to the rejection of the null hypothesis is $1 - (1 - 0.05)^{10} = 40\%$, even if H_0 is always true. This measures the overall probability of a false positive among all overbidding tests. A standard, but rather conservative, procedure to address this issue is the Bonferroni correction. If h tests are performed with the same sample, the significance value α must be reduced to $m = \alpha/h$. For a series of 10 tests, the significance value after the Bonferroni correction falls to 0.005, and the probability of obtaining at least one significant result is therefore 4.89%. A slightly different procedure, the Sidak correction, sets m such that the probability of obtaining at least one significant result remains equal to 5%. The Bonferroni procedure is conservative as it assumes independence between each pair in the series of tests, which is not always the case. The series of 10 bids elicited from the same subjects are likely correlated, so that independence does not hold. In this case, the significance level set by the Bonferroni correction leads to an increase in type II errors. In other words, the Bonferroni correction reduces the risk of false positives at the cost of a greater risk of false negatives. When the correlation structure between the tests is taken into account, the significance level lies between α and α/h. However, if the experiment was not intended *ex ante* to elicit 10 bids but only to test the value of the last one, to pick up overbidding once learning has occurred, for example, there is no reason to apply the Bonferroni correction for this one test. Step-up procedures are more powerful than the Bonferroni correction for multiple comparisons. These should be preferred if false negatives are an issue. A step-up procedure ranks each hypothesis according to its p-value, and then assigns different significance levels according to this rank. One popular step-up alternative to the Bonferroni correction is the Holm–Bonferroni procedure. The decision rule in the Holm–Bonferroni procedure is to not reject those hypotheses for which the rank-ordered p-value $p_{(k)}$ is larger than $\frac{\alpha}{h+1-k}$. The Westfall–Young permutation procedure combines the Holm–Bonferroni procedure with bootstrapping to calculate the sampling distribution of the p-values. Shaffer (1995) provides an accessible review of multiple-testing procedures. The Benjamini–Hochberg procedure requires the expected rate of false positives to be under α. When the h tests are independent, all hypotheses with a rank-ordered p-value $p_{(k)}$ under $\frac{k\alpha}{h}$ are rejected.

into account the multiplicity of hypothesis testing – Focus 7.8 describes several such procedures.

The probability of a type II error is

$$\beta(\theta) = Pr[d_0|H_1] = Pr[\mathbf{T}(\mathbf{Y}) \notin R|\theta \in \Theta_1]$$

The statistical power of a test is the probability of making a correct decision, i.e. of rejecting the null hypothesis when the true parameter corresponds to the alternative hypothesis. This is defined as:

$$Pr[\mathbf{T}(\mathbf{Y}) \in R|\theta \in \Theta_1] = 1 - \beta(\theta)$$

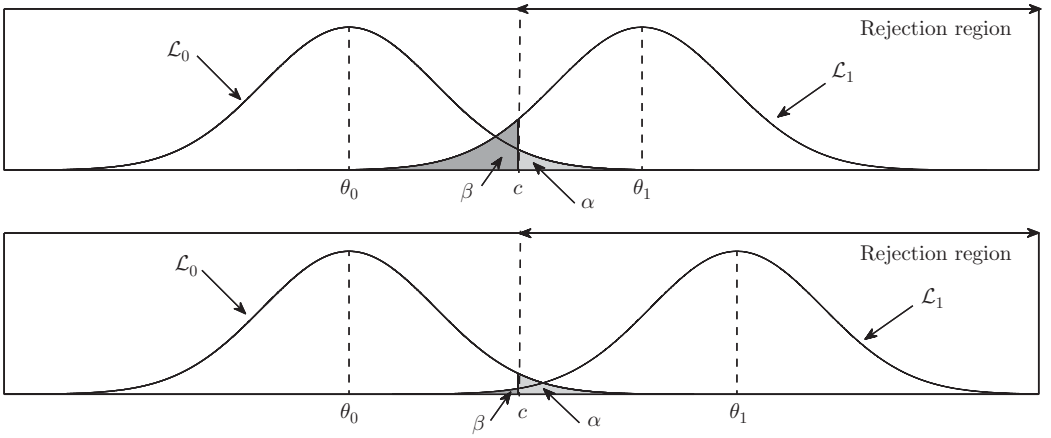


Figure 7.12 Power under different alternative hypotheses

The power is not constant and depends on which alternative hypothesis represents the true DGP. In other words, power is not a number but a function of the real value of θ under the alternative hypothesis. Figure 7.12 shows how the power $1-\beta$ changes as the value of θ under the alternative hypothesis moves. In the top panel, θ_1 is close to θ_0 and the probability of a false negative is high: the test has low power. In the left panel, the value of θ_1 increases and the probability of a false negative falls. The latter test has higher power than the former.

Example: Consider a sample $\mathbf{Y}$ *such that* $Y_i \sim \mathcal{N}(\theta, \sigma^2)$*, and let* $\hat{\theta} = \bar{Y}$ *be an estimator of* θ*. We want to test* $H_0 : \theta = 2$ *against* $H_1 : \theta > 2$*. The rejection region R is such that* $\bar{Y}$ *must be large enough to lead to decision* d_1*. With a significance level* α*, the probability of observing* $\bar{Y}$ *in the rejection region R when* H_0 *corresponds to the true DGP is* $Pr[\frac{\bar{Y}-2}{\sigma/\sqrt{n}} > c] = \alpha$*.*

- *The rejection region R is such that the sample value of the test statistic* $\mathbf{T}(\mathbf{Y}) = \frac{\bar{Y}-2}{\sigma/\sqrt{n}}$ *is larger than the critical value* $c = z_\alpha$*. Decision* d_0 *is taken if the sample value of* $\mathbf{T}(\mathbf{Y})$ *is lower than this critical value.*
- *If* $\alpha = 0.05$*,* $n = 64$*,* $\sigma = 4$ *and* $\bar{y} = 3$*, the test statistic is* $\mathbf{T}(\mathbf{y}) = (3-2)/(4/\sqrt{64}) = 2$ *and the critical value is* $z_{0.05} = 1.645$*. The decision rule is to take decision* d_1 *and reject hypothesis* H_0*. The probability of observing the sample mean under hypothesis* H_0 *is 0.023.*
- *When the alternative hypothesis is* $H_1 : \theta > 4$*, the power of the test is* $1 - Pr[\frac{\bar{Y}-4}{\sigma/\sqrt{n}} < z_\alpha]$*. When the alternative hypothesis is* $H_1 : \theta > 10$*, the power of the test is much larger and equals* $1 - Pr[\frac{\bar{Y}-10}{\sigma/\sqrt{n}} < z_\alpha]$*. This illustrates that the power of the test increases with the size of the gap between the null and alternative hypotheses. Figure 7.12 also illustrates this point.*
- *Suppose that* $n = 6,400$ *instead of* $n = 64$ *and* $\bar{y} = 2.01$*. In this case, the test statistic is* $\mathbf{T}(\mathbf{y}) = (2.1-2)/(4/\sqrt{6,400}) = 2 > z_{0.05}$*. With very large samples, any*

Focus 7.9
Sample-size determination

In hypothesis testing, the size of the sample n is taken as given. Sample-size determination allows the sample size to be chosen in order to achieve a size α and a power $1-\beta$, given the null and alternative hypotheses. The sample size is determined by solving the following two equations:

$$\text{Size} = Pr[\mathbf{T}(\mathbf{Y}) \in R | \theta \in \Theta_0] \leq \alpha$$
$$\text{Power} = Pr[\mathbf{T}(\mathbf{Y}) \in R | \theta \in \Theta_1] \geq 1-\beta$$

Suppose that we would like to test $H_0 : \theta = 2$ against $H_1 : \theta = 3$ with a sample from a normally distributed population with mean θ and a sample standard deviation of 3. How many subjects would we need to carry out a test with significance level $\alpha = 0.05$ and power $1-\beta = 0.90$? Solving the equations yields a minimum sample size of 79 subjects required for this test.

small difference produces statistical significance. This often leads to distinguishing statistical significance, which is a formal result, and practical significance, in which significance is put into perspective with the size of the sample used to obtain the significant difference.

The probability of making a correct decision depends on four variables: the sample size used to make the decision, the level of significance (the type I error), the rejection region R, and the size of the gap between the null and alternative hypotheses. The ideal test would combine a low type I error (a low rate of false positives) and a high power (a low rate of false negatives). There is, however, a trade-off between type-I and type II errors. As illustrated in Figure 7.12, any increase in type I error α comes at the price of a decrease in type II error β; and any decrease in type I error results in an increase in type II error. As a consequence, both risks cannot be minimised at the same time. The Neyman principle solves the trade-off by arbitrarily choosing the size α. The test is then selected based on the highest power under H_1. If such a test exists, it is the most powerful test; if the power, moreover, approaches 1 as the sample size goes to infinity, the test is said to be consistent.

An alternative to full hypothesis testing is to calculate a *p-value* or significance probability. A p-value is defined as the probability of observing a test statistic that is as extreme as the observed test statistic $\mathbf{T}(\mathbf{y})$ (e.g. $\mathbf{T}(\mathbf{Y}) > \mathbf{T}(\mathbf{y})$ or $\mathbf{T}(\mathbf{Y}) < \mathbf{T}(\mathbf{y})$, or both), when the DGP is that in the null hypothesis. A p-value is the lowest significance level for which the null hypothesis should be rejected. A small p-value is taken as evidence against the null hypothesis. Focus 7.9 shows how sample size and p-values can be related.

Significance tests should, however, be used with caution as they do not allow researchers to state evidence in favour of the null hypothesis, and they tend to overstate the evidence against the null hypothesis. First, a p-value does not answer the question 'How probable is the validity of the null hypothesis, given the sample data?' but rather

'How probable are the sampled data, given that the null hypothesis is true?' The point is illustrated by Rouder et al. (2009) using the following example. Consider a simple one-sample *t*-test of whether a population mean is different from 0. If the null hypothesis is false (the population mean actually is different from 0), the values of the *t*-statistic increase without bound and the *p*-values for the *t*-test converge on 0 as the sample size rises. In other words, the significance test has the desirable property of rejecting the null hypothesis as more and more data are collected. If, however, the null hypothesis is true (the population mean is actually 0) and the sample size is over 30, the sampling distribution of the *t*-statistic can be approximated by a standard normal distribution and all *p*-values are equally likely. In this case, increasing the sample size does not increase the evidence in favour of the null hypothesis. In significance tests, a null hypothesis can thus only be rejected or not rejected, and can never be 'accepted' or confirmed by the empirical evidence. Second, the *p*-value is itself a random variable. For example, the *p*-value is generally uniformly distributed between 0 and 1 under the null hypothesis.[13] As a result, the *p*-value can be seen as a random draw from the unit interval if H_0 is true, while it is concentrated closer to 0 if, rather, H_1 is true. A small *p*-value can thus result by chance, even if computed on data produced by H_1. Third, a large *p*-value can reflect an insufficiently large sample or the use of an inappropriate test statistic. An alternative way of evaluating evidence in hypothesis testing is the Bayes factor, described in Focus 7.10.

7.3.2 Non-parametric versus Parametric Tests

Parametric tests are appropriate when the population distribution can be assumed to be approximatively normal. This especially applies to large samples, thanks to the central limit theorem. Parametric tests are preferred in two kinds of application. First, non-parametric tests are well suited to nominal and ordinal scales as there are no parametric methods for this kind of estimator. Second, they are also well suited to interval and ratio scales when the population distribution is unspecified or cannot be approximated by a normal distribution. Siegel (1957) suggests that the choice between parametric and non-parametric tests should be based on three criteria:

C1: The applicability of the statistical models on which the tests are based to the observed data. For example, most parametric tests assume that the variables are measured on interval or ratio scales, that the observations are independent, are drawn from a normally distributed population, and have the same variance. The statistical models underlying non-parametric tests are, in general, less restrictive.[14] With small samples, non-parametric tests are the only possible alternative, unless some prior knowledge about the population distribution is available.

[13] This applies when the alternative hypothesis is simple, the distribution is continuous and the test is of the Neyman–Pearson type.

[14] Parametric tests are not always more demanding in terms of assumptions than non-parametric tests. For example, comparing the variances of two populations requires fewer assumptions via a Fisher test than via its non-parametric counterpart.

Focus 7.10
Bayes factors

Bayes factors avoid the two key criticisms against significance tests. In particular, Bayes factors produce statements about the likelihood of the null hypothesis. The principle is based on Bayesian statistics: any prior opinion is transformed into a posterior opinion through consideration of the data. In this setting, the data Y are assumed to have arisen from either one hypothesis (H_0) or the other (H_1). The researcher has priors for each hypothesis $Pr[H_0]$ and $Pr[H_1]$. Under hypothesis H_0, the probability density for observing data Y is $Pr[Y|H_0]$. Under hypothesis H_1, it is $Pr[Y|H_1]$. By Bayes's theorem, the posterior probability that hypothesis $H_k, k = 1, 2$ is true given data Y is

$$Pr[H_k|Y] = \frac{Pr[Y|H_k] \times Pr[H_k]}{Pr[Y]}$$

and the ratio of the two probabilities is:

$$\underbrace{\frac{Pr[H_0|Y]}{Pr[H_1|Y]}}_{\text{posterior odds}} = \underbrace{\frac{Pr[Y|H_0]}{Pr[Y|H_1]}}_{\text{Bayes factor}} \times \underbrace{\frac{Pr[H_0]}{Pr[H_1]}}_{\text{prior odds}}$$

In other words, the Bayes factor is the ratio of the posterior and prior odds of H_0. Bayes factors are usually interpreted using the following nomenclature:

B_{10}	Evidence against H_0
1–3	not worth more than a bare mention
3–20	substantial
20–150	strong
>150	decisive

The computation of the factor is application-specific. A number of techniques are summarised in e.g. the review by Kass and Raftery (1995).

C2: The level of measurement, i.e. the measurement scale. Parametric tests can only be used for interval and ratio scales. For nominal or ordinal scales, like ranks, scores or classifications, only non-parametric tests can be used.

C3: The power efficiency of the alternative tests. Due to the strength of their assumptions, parametric tests are more powerful than non-parametric tests.

Table 7.5 shows how to choose among the most frequently used statistical tests, according to the level of measurement, the assumptions about the population distribution (if any) and the characteristics of the sample(s). Focus 7.11 shows a very general approach to hypothesis testing with likelihood-ratio tests.

7.3.3 One-Sample Tests

This section focuses on one-sample tests that can be used to decide on the sample from a specific population. This principle can be used to improve the informational content of

Table 7.5 Frequently used statistical tests

	Level of measurement and parametric assumptions		
	Interval/ratio and normal	Ordinal or interval/ratio and not normal	Categorical
One sample			
	t-test	Wilcoxon test	Binomial test
	z-test	Sign test	
	Chi-square test for the variance	Kolmogorov–Smirnov test	Chi-squared test
Independent samples			
2-sample	*t*-test	Mann–Whitney test	Fisher exact test
	z-test	Kolmogorov–Smirnov test	Chi-squared test
	Welch's test		
	F-test	Siegel–Tukey test	
K-sample	One-way ANOVA	Kruskal-Wallis test	Chi-squared test
	Barlett's test	Levene's test	
Dependent samples			
2-sample	Paired *t*-test	Matched-pairs Wilcoxon test	McNemar test
K-sample	Repeated-measure ANOVA	Friedman test	Cochran's Q test

data, by assessing whether observations are drawn in a common distribution – a typical example is statistical tests used to detect outliers, presented in Focus 7.12. We present three main types of one-sample test, depending on the level of measurement. If the data are interval (or ratio) and normal, the one-sample tests are parametric.[15] If the data are ordinal or interval but not normally distributed, the one-sample tests are non-parametric. Last, we present one-sample tests when the data are categorical. In what follows, the sample variables Y_i are supposed to have been independently and identically drawn from the same population distribution. Unless specified otherwise, the hypothesis has the form: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

Interval and Normal Variables

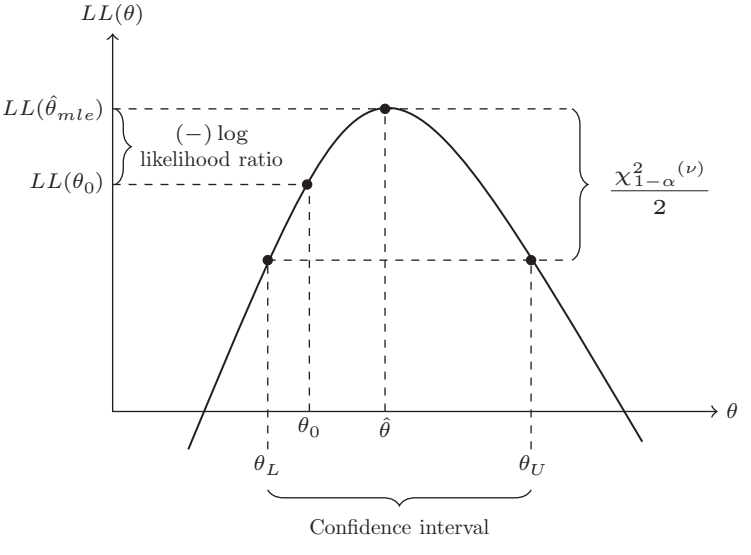
The value of the mean: t-test and z-test

One of the simplest hypothesis tests is the one-sample *t*-test. This can be used to test if a particular sample of participants in an experiment is similar or differs from a benchmark. In economics, benchmarks are either set to predetermined theoretical values or based on stylised facts. The estimator of the mean is $\hat{\theta} = \bar{Y}$ and the test statistic is

[15] For the sake of simplicity, we refer to the ratio and interval levels of measurement as 'interval' in the following. The data are supposed to be drawn from continuous and strictly increasing population distributions.

Focus 7.11
The likelihood-ratio test

Likelihood-ratio tests provide a very general approach to hypothesis testing. The logic of the test is the following. Suppose that the data $(y_1, \ldots, y_n)$ are generated by a probability-density function $g(y_i|\theta)$ with parameter θ. The likelihood of the random sample $\mathbf{Y}$ for a given parameter value θ is $L(\theta) = \prod_{i=1}^{n} g(y_i|\theta)$. Suppose the hypothesis test has the form $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Under the null hypothesis H_0, the value of the likelihood at $\theta = \theta_0$ is expected to be relatively large. To judge whether this likelihood value is large, we need a reference point. This reference point is the maximum value of the likelihood, attained at value $\hat{\theta}_{mle}$: $L(\hat{\theta}_{mle}) = max_{\theta} L(\theta)$. As the latter is the maximum likelihood, it is always greater than or equal to the likelihood $L(\theta_0)$ evaluated at the base hypothesis. The likelihood-ratio test compares these two likelihood values using their ratio: $L_r = \frac{L(\theta_0)}{L(\hat{\theta}_{mle})}$. The likelihood ratio is between 0 and 1. If the likelihood ratio is small, there is a large discrepancy between the parameter that actually maximises the likelihood and θ_0 – which leads to doubting the validity of the base hypothesis. Under regularity conditions and under H_0, the large sample distribution of $-2 \log L_r$ is chi-squared. The degrees of freedom are the number of restrictions imposed in the base hypothesis (in our example, the degree of freedom is one). The likelihood-ratio test cannot be used to compare two simple hypotheses (e.g. $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$) as in this case the degree of freedom of the chi-squared distribution is 0 and the test statistic is undefined. Note that using the chi-squared distribution as the limiting distribution of $-2 \log L_r$ hinges critically on the regularity conditions for the density function $g(y_i|\theta)$. These regularity conditions are set out in, e.g., Wooldridge (2002) and Davidson and MacKinnon (2004). The figure below illustrates the logic of the likelihood-ratio test.



The curve depicts the logarithm of $L(\theta)$ for different values of θ. The maximum value of the log likelihood $LL(\theta)$ is attained at $\theta = \hat{\theta}_{mle}$. The logarithm of the likelihood ratio corresponds to the distance between the maximum value of the log likelihood, $LL(\hat{\theta}_{mle})$, and the log likelihood evaluated at θ_0, $LL(\theta_0)$. The figure also shows how to construct a confidence interval for a (log) likelihood ratio. For a given value of the chi-squared statistic $\chi^2_{1-\alpha}(\nu)$ with ν degrees of freedom, corresponding to the number of restrictions under H_0, the log likelihood curve defines two bounds θ_L and θ_U such that the distance between the maximum log likelihood and the log-likelihood evaluated at H_0 is no greater than one-half of this chi-squared value (noting that the large-sample distribution of $-2 \log L_r$ is chi-squared).

$\mathbf{T}(\mathbf{Y}) = \frac{\bar{Y}-\theta}{S/\sqrt{n}}$. For a sample of variables Y_i, normally distributed with unknown variance σ^2, the sampling distribution of the test statistic under the null hypothesis is a t-distribution with $n-1$ degrees of freedom: $\frac{\bar{Y}-\theta_0}{S/\sqrt{n}} \sim St(n-1)$, and the rejection region is such that:

$$c_l = -t_{\frac{\alpha}{2}} \text{ and } c_u = t_{\frac{\alpha}{2}} :$$

with the t-score having $n-1$ degrees of freedom. In the one-sided test, the rejection region is such that $c = t_\alpha$ (for $H_1 : \theta > \theta_0$) or $c = -t_\alpha$ (for $H_1 : \theta < \theta_0$). If the variance σ^2 is known, the critical value $\mathbf{T}(\mathbf{Y})$ is replaced by the corresponding z-score, and the sample variance S^2 is replaced by σ^2. In this case, the test is called a z-test. When the sample size is very large, the z-test and the t-test are asymptotically equivalent.

Example: Consider an experiment in which 54 participants bid in second-price auctions. The expected payoff in the auction is \$10 for rational bidders. The data can be used to answer the question 'Due to overbidding, is the observed average payoff significantly lower than expected?', by testing the following hypotheses:

$$\begin{cases} H_0 : & \textit{The average payoff corresponds to rational behaviour, } \theta = 10; \\ H_1 : & \textit{The average payoff is lower than under rational behaviour, } \theta < 10. \end{cases}$$

Under the assumption that the sample distribution is approximatively normal, the t-test provides a statistical answer to this question. Suppose the sample mean is $\bar{y} = 8.89$ *and the sample standard deviation is* $s = 2.76$. *The test statistic is* $\mathbf{T}(\mathbf{Y}) = \frac{\bar{y}-\theta_0}{s/\sqrt{n}} = \frac{8.89-10}{2.76/\sqrt{54}} = -2.96$ *and the sampling distribution is* $\frac{\bar{Y}-\theta_0}{S/\sqrt{n}} \sim St(53)$. *The critical value at* $\alpha = 0.05$ *is* $c = t_{0.025} = -2.006$ *for a t-score with 53 degrees of freedom. According to the decision rule, the hypothesis* H_0 *is rejected and the mean observed payoffs are not consistent with rational behaviour.*

The value of the variance: Chi-squared test for the variance

This test can be used to assess if the dispersion in a particular sample of participants differs from some given level. The estimator of the unknown variance θ^2 is the sample variance $S^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$ and the test statistic is $\mathbf{T}(\mathbf{Y}) = \frac{(n-1)S^2}{\theta^2}$. The sampling distribution under the null hypothesis is $\frac{(n-1)S^2}{\theta_0^2} \sim \chi^2(n-1)$. The rejection region is

$$c_l = \chi^2_{n-1,1-\frac{\alpha}{2}} \text{ and } c_u = \chi^2_{n-1,\frac{\alpha}{2}}$$

Focus 7.12
Testing for outliers

A number of parametric tests can be used to detect outliers. They all assume that the underlying sample (minus any outlier) is normally distributed. The **Dixon test** is used to detect a single outlier in a set of observations drawn from a normal distribution when the sample size is over three and under 25. The null hypothesis is that there is no outlier. The test statistic is based on a rank-ordering of observations $Y_{(1)} \leq \ldots \leq Y_{(n)}$. The test statistic $\mathbf{T}(\mathbf{Y})$ is equal to the difference between the suspect value and its nearest neighbour (for $n \leq 10$) or its second-nearest neighbour (for $11 \leq n \leq 25$), divided by the range of the observations. The following table shows the value of the test statistic depending on the number of observations and the position of the possible outlier:

Sample size	High outlier $Y_{(n)}$	Low outlier $Y_{(1)}$
$3 \leq n \leq 7$	$\frac{Y_{(n)}-Y_{(n-1)}}{Y_{(n)}-Y_{(1)}}$	$\frac{Y_{(2)}-Y_{(1)}}{Y_{(n)}-Y_{(1)}}$
$8 \leq n \leq 10$	$\frac{Y_{(n)}-Y_{(n-1)}}{Y_{(n)}-Y_{(2)}}$	$\frac{Y_{(2)}-Y_{(1)}}{Y_{(n-1)}-Y_{(1)}}$
$11 \leq n \leq 13$	$\frac{Y_{(n)}-Y_{(n-2)}}{Y_{(n)}-Y_{(2)}}$	$\frac{Y_{(3)}-Y_{(1)}}{Y_{(n-1)}-Y_{(1)}}$
$14 \leq n \leq 25$	$\frac{Y_{(n)}-Y_{(n-2)}}{Y_{(n)}-Y_{(3)}}$	$\frac{Y_{(3)}-Y_{(1)}}{Y_{(n-2)}-Y_{(1)}}$

The sample value of the test statistic is compared to a critical value. The requirements of the two underlying assumptions (a normal distribution and small sample size) show that the Dixon test should be used with caution. Moreover, masking can occur if there is more than one outlier in the data. In this case, one outlier masks another as the distance between the two outliers (e.g. the distance between $Y_{(n)}$ and $Y_{(n-1)}$) is not sufficient to indicate a significant difference. The **Grubb test** is an alternative for the detection of a single outlier in a set of observations. The test statistic is the largest standardised absolute deviation from the sample mean. The test statistic is $\mathbf{T}(\mathbf{Y}) = \frac{max|Y_i-\bar{Y}|}{S}$ and the rejection region is such that $\mathbf{T}(\mathbf{Y}|H_0) > c$, with:

$$c = \frac{n-1}{n}\sqrt{\frac{(t_{\alpha/(2n)})^2}{n-2+(t_{\alpha/(2n)})^2}}$$

where t is the t-score with $n-2$ degrees of freedom. Both the Dixon and Grubb tests detect single outliers. If more than one outlier is suspected, multiple-testing issues and 'masking' can arise. For large sample sizes ($n > 20$) and a normal population distribution, the **Rosner test** allows the detection of up to 10 outliers in a sample. The test assumes that the outliers come from a different distribution from the normal population distribution. For h outliers to be detected, the test uses an iterative series of r sub-tests. First, the observation $Y^{(0)}$ that is furthest from the sample mean is identified and used to calculate the test statistic $\mathbf{T}(\mathbf{Y}^{(0)}) = \frac{|Y^{(0)}-\bar{Y}|}{S}$. Then observation $Y^{(0)}$ is deleted from the sample, and the new values of the sample mean $\bar{Y}^{(1)}$ and standard deviation $S^{(1)}$ are calculated. Again, the observation $Y^{(1)}$ furthest from the sample mean $\bar{Y}^{(1)}$ is identified, and the test statistic $\mathbf{T}[\mathbf{Y}^{(1)}] = \frac{|Y^{(1)}-\bar{Y}^{(1)}|}{S^{(1)}}$ is calculated. Observation $Y^{(1)}$ is then removed and a new test statistic $\mathbf{T}[\mathbf{Y}^{(2)}]$ is calculated.

The process continues until h outliers have been removed from the data. If the test statistic $\mathbf{T}[\mathbf{Y}^{(h-1)}]$ is over the critical value, then h outliers are detected; if it is under the critical value, then the next test statistic $\mathbf{T}[\mathbf{Y}^{(h-2)}]$ is used to test for $h-1$ outliers. The procedure continues until a certain number of outliers have been detected. If none of the test statistics are significant, the null hypothesis of the absence of outliers in the data cannot be rejected. While the Rosner test is immune to 'masking', it is not immune to 'swamping', where the test detects a block of outliers of which only a certain number are actual outliers.

With a one-sided test, the lower rejection region is $c = \chi^2_{n-1,1-\alpha}$ and the upper rejection region is $c = \chi^2_{n-1,\alpha}$. If the mean of the population distribution μ is known, the critical values are taken from a chi-squared distribution with n degrees of freedom.

Example: Consider again the above-mentioned experiment on second-price auctions. Based on the available evidence the payoff distribution is expected to have a standard deviation of 1$. The hypotheses associated with the question 'Is the observed payoff dispersion greater than this?' are:

$$\begin{cases} H_0: & \textit{The standard deviation is } \theta = 1; \\ H_1: & \textit{The standard deviation is larger, } \theta > 1. \end{cases}$$

The chi-squared test allows us to answer this question. The sample standard deviation is $s = 2.7$. *The test statistic is* $\mathbf{T}(\mathbf{Y}) = \frac{(n-1)s^2}{\theta_0^2} = \frac{(54-1)*(2.7)^2}{1^2} \simeq 402.8$. *The sampling distribution is* $\frac{(n-1)S^2}{\theta_0^2} \sim \chi^2(53)$. *The critical value at* $\alpha = 0.05$ *is* $c = \chi^2_{53,0.95} \simeq 71$. *According to the decision rule, hypothesis* H_0 *is rejected and the dispersion inferred from the observed payoffs is greater than the benchmark value of 1.*

Interval and Ordinal Variables

Non-parametric tests can be used when the data are measured on an ordinal scale, or are interval variables for which normality cannot be assumed. The main one-sample non-parametric test applies to the median. One advantage of the median as the central tendency parameter of a distribution is that it always exists, which is not the case for the mean. As discussed in Section 7.1.4, the median is also a more robust estimate of the central tendency. The median is not, however, the only quantile for which non-parametric tests can be carried out. Gibbons (2011) sets out the procedures for hypothesis tests regarding any specified quantile. When the data are measured on the interval scale, the observed sample $\mathbf{y} = (y_1, \ldots, y_n)$ is supposed to be drawn, independently and identically, from a continuous and strictly increasing population distribution.

Sign test

The sign test refers to the value of an unknown median θ of the population. The logic behind the sign test is to dichotomise the sample observations: one part below the median and the other above. The test statistic is based on counts, i.e. on the number of plus/minus signs among the n differences $Y_i - \theta, i = 1, \ldots, n$. The test statistic is defined

as the max between $\mathbf{T}^+$ and $\mathbf{T}^-$, where $\mathbf{T}^+$ is the number of observations above θ and $\mathbf{T}^-$ the number below θ. The sampling distribution of $\mathbf{T}(\mathbf{Y})$ under the null hypothesis is a binomial distribution $\mathrm{B}(n, 0.5)$. The rejection region is such that:

$$\begin{array}{ll} c_l & \text{is the largest integer such that } \sum_{k=0}^{c_l} \binom{n}{k} 0.5^n \leq \frac{\alpha}{2} \\ c_u & \text{is the smallest integer such that } \sum_{k=c_u}^{n} \binom{n}{k} 0.5^n \leq \frac{\alpha}{2} \end{array}$$

Example: In the above-mentioned experiment on second-price auctions, imagine you no longer want to assume that the population distribution is approximatively normal. You expect the median payoff to be \$10, the figure expected for a rational individual, and test the hypotheses:

$$\begin{cases} H_0: & \textit{The average payoff corresponds to rational behaviour, } \theta = 10; \\ H_1: & \textit{The average payoff is lower than under rational behaviour, } \theta < 10. \end{cases}$$

The sign test can be applied to answer the question. Say the statistic is 31. The critical values at $\alpha = 0.05$ *are* $c_l = 20$ *and* $c_u = 32$*. According to the decision rule, the hypothesis* H_0 *is not rejected and the median observed payoff is consistent with rational behaviour.*

Zero differences between Y_i and θ_0 are usually ignored and the sample size is reduced accordingly.[16] For a one-sided test, the test statistic is either $\mathbf{T}^+$, the number of observations greater than θ (if $H_1 = \theta > \theta_0$), or $\mathbf{T}^-$, the number of observations less than θ (if $H_1 = \theta < \theta_0$). The sign test is a special case of the binomial test where the probability of success under the null hypothesis is $p = 1/2$. For $n > 25$ a normal approximation can be used. As the normal approximation works very well for binomial distributions with parameter one-half, this approximation is fairly robust to small samples. In this case, the test statistic is $\frac{\mathbf{T}(\mathbf{Y}) - 0.5 \times n}{0.5\sqrt{n}}$ and the critical values are the usual z-scores.

Wilcoxon test

The Wilcoxon test is the main non-parametric test regarding the value of an unknown median θ in the population. The test is more powerful than the sign test. As is usual in hypothesis testing, increased power comes at the price of additional assumptions. The Wilcoxon test assumes that the population distribution is symmetric around θ. As such, the median equals the mean and inference about one of the two also holds for the other. The Wilcoxon test is often called the Wilcoxon signed-rank test, as the test statistic combines the ranks of differences and their signs. The test statistic is based on the ranking of absolute differences $|\Delta_i| = |Y_i - \theta|$, ranked from the smallest to the largest. The rank assigned to a given difference Δ_i is denoted $r(\Delta_i)$. The test statistic is the expected value of the sum of the ranks for positive ($\mathbf{T}^+$) and negative ($\mathbf{T}^-$) differences $\Delta_i = Y_i - \theta$, as follows:

$$\mathbf{T}^+ = \sum_i Z_i r(|\Delta_i|) \text{ and } \mathbf{T}^- = \sum_i (1 - Z_i) r(|\Delta_i|)$$

with $Z_i = 1$ if $\Delta_i > 0$ and zero otherwise, and where $r(\Delta_i)$ is the rank of absolute differences $|\Delta_i|$ ranked from the smallest to the largest. Under H_0, symmetry implies

[16] There are alternative approaches, e.g. to treat half the zeros as pluses and the other half as minuses, to assign a sign at random to zeros, and to assign to all of the zeros the sign which is most in favour of H_0.

that $\mathbf{T}^+ = \mathbf{T}^-$ and the sampling distribution of $\mathbf{T}^+$ is $Pr[\mathbf{T}^+ = h] = u(h)/2^n$, where $u(h)$ is the number of possible ways of assigning plus and minus signs to the first n integers such that the sum of the positive integers equals h. The rejection region is

$$R > c \text{ or } R < c, \text{ with } c \text{ such that } Pr[\mathbf{T}^+ > c|H_0] = \alpha$$

For a sample size $n > 20$ a normal approximation can be used. In this case, the test statistic is $\frac{4\mathbf{T}^+ - n(n+1)}{\sqrt{2n(n+1)(2n+1)/3}}$ and the critical values are the corresponding z-scores.

Example: In the above-mentioned experiment on second-price auctions, the Wilcoxon test allows us to test the hypotheses that payoffs match those of rational players if the population distribution is no longer assumed to be normally distributed, but is nevertheless assumed to be symmetric.

$$\begin{cases} H_0: & \textit{The average payoff corresponds to rational behaviour, } \theta = 10; \\ H_1: & \textit{The average payoff is lower than under rational behaviour, } \theta < 10. \end{cases}$$

The test statistic is 341.5, with a p-value equal to 0.007. According to the decision rule, the hypothesis H_0 is rejected and the median observed payoff is not consistent with rational behaviour.

The Wilcoxon test is a special case of the more general class of rank-order statistics. For a sample of observations $\mathbf{y} = (y_1, \ldots, y_n)$, a rank-order statistic $r(\mathbf{y}) = [r(y_1), \ldots, r(y_n)]$ is any function such that $r(y_i) \leq r(y_j)$ when $y_i \leq y_j$. The continuity assumption made for the population distribution is important as it implies (theoretically) that the probability of two observations with identical values is 0, as such a sample with n observations generates n ranks. This raises a practical difficulty when there is a difference of 0 (when observations Y_i are equal to θ_0), which contradicts continuity. One solution is to ignore the 0 difference and reduce the sample size accordingly. With more than one 0 difference, another possibility is to correct for these ties by assigning each 0 the average of the ranks they would have if they were ranked as plus or minus differences. Gibbons (2011) discusses a number of alternative methods to correct for ties. Even so, data with numerous ties remain problematic with the Wilcoxon test.

One criticism of the Wilcoxon test is that the null hypothesis is in fact composite, as it assumes that the distribution is symmetric and has median θ_0. If the null hypothesis is rejected, then either the distribution is not symmetric, or it is symmetric but has a median different from θ_0.

Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test is a test of the goodness of fit. Compared to the Wilcoxon test, the Kolmogorov–Smirnov test focuses on the shape of the distribution and not only on its location. The Kolmogorov–Smirnov test compares the counts of a univariate variable with the expected counts from the assumed continuous distribution. The test statistic is based on the differences between the empirical distribution function of the sample G_o (i.e. the proportion of sample observations that are less than or equal to a value a, for all a) and an assumed cumulative distribution function G_e. The Kolmogorov–Smirnov test is formalised by setting up the hypothesis in terms of the equality of distributions:

$$H_0 : G_o(Y_i) = G_e(Y_i), \forall i \text{ vs. } H_1 : G_o(Y_i) \neq G_e(Y_i) \text{ for some } i$$

The test statistic compares the two cumulative distribution functions (the empirical and the assumed) and focuses on the point of maximum difference between the two distributions. This is defined as $\mathbf{T}(\mathbf{Y}) = sup_y|G_o(y) - G_e(y)|$. If the sample comes from the theoretical distribution, the test statistic tends to a maximum difference of 0 as the sample size rises. If, on the contrary, the two distributions do not overlap, the maximum difference is 1. Under H_0, the sampling distribution of $\mathbf{T}(\mathbf{Y})$ is independent of G_e. As the description of the sampling distribution is tedious, we do not present it here. The rejection region is $R > c$ such that $Pr[\mathbf{T}(\mathbf{Y}) > c|H_0] = \alpha$.

The theoretical properties of the Kolmogorov–Smirnov test are only valid when G_e is assumed to be continuous. In practice, the test can be extended to discrete distributions. The Kolmogorov–Smirnov test requires a relatively large sample size in order to mimic a continuous distribution and reject the null hypothesis.[17] When G_e is a normal distribution with unspecified parameters, the test becomes the Lilliefors test. Focus 7.13 contains more specialised tests for normality and Focus 7.14 presents a test for randomness.

Example: Consider the test of the normality of the population distribution of payoffs, $\mathcal{L}$, in the above-mentioned experiment on second-price auctions,

$$\begin{cases} H_0 : & \textit{The payoffs are normally distributed, } \mathcal{L} \equiv \mathcal{N}(10, 2); \\ H_1 : & \textit{The payoffs are not normally distributed, } \mathcal{L} \neq \mathcal{N}(10, 2). \end{cases}$$

Such hypotheses can be tested using a Kolmogorov–Smirnov test. The sample value of the test statistic is 0.2647, with a p-value of 0.001. According to the decision rule, hypothesis H_0 is rejected and the payoffs are not normally distributed $\mathcal{N}(10, 2)$.

Categorical Variables

Binomial test

When there are only two categories, the binomial test is the first way of comparing the observed distribution G_o to an expected distribution G_e. The sample consists of categorical variables $Y_i = \{0, 1\}$ in unknown proportions $1 - \theta$ and θ. The estimator is the sample frequency $\hat{\theta} = \bar{y}$ and the test statistic is the number of successes in the Bernoulli trials $\mathbf{T}(\mathbf{Y}) = \sum_i Y_i$. Under the null hypothesis, the test statistic follows a binomial distribution $\mathrm{B}(n, \theta_0)$ and the rejection region is such that:

c_l is the largest integer such that $\sum_{i=0}^{c_l} \binom{n}{i} \theta_0^i (1 - \theta_0)^{n-i} \leq \frac{\alpha}{2}$
c_u is the smallest integer such that $\sum_{i=c_u}^{n} \binom{n}{i} \theta_0^i (1 - \theta_0)^{n-i} \leq \frac{\alpha}{2}$

Example: Consider an experiment in which 19 participants over 54 are female. You want to answer the question, 'Is the observed proportion of 19/54 significantly lower than the figure that would be expected?', i.e. *is the subject pool random as regards gender:*

$$\begin{cases} H_0 : & \textit{The proportion of female participants is } \theta = 0.5; \\ H_1 : & \textit{The proportion of female participants is } \theta < 0.5. \end{cases}$$

[17] If the sample size is too small, not enough points are used to construct a satisfactory empirical distribution function.

Focus 7.13
Goodness-of-fit tests and the normality hypothesis

Goodness-of-fit tests provide inference about the form of the population from which the sample was drawn, for instance whether or not the population distribution is normal. As most of the parametric tests assume normality, such tests are particularly useful in deciding whether to use parametric or non-parametric tests. A goodness-of-fit test of normality is defined as:

$$\begin{cases} H_0: & \text{The data come from draws in a normally distributed population;} \\ H_1: & \text{The data do not come from draws in a normally distributed population} \end{cases}$$

If the null hypothesis refers to a discrete distribution, the goodness-of-fit test is chi-squared, as described in the main text. The test statistic is the sum of squares of the gaps between the observed and expected frequencies in each of the K classes of the discrete distribution, normalised by the expected frequency. The statistic is distributed approximately chi-squared with $K-1$ degrees of freedom. If the null hypothesis rather refers to a continuous hypothesis, one of the following tests can be applied.

- The **Kolmogorov–Smirnov test** is described for ordinal data in the main text, and can be used to test for goodness of fit with a given cumulative distribution function. The test statistic is based on the differences between the empirical distribution function of the sample G_o (i.e. the proportion of sample observations that are less than or equal to a value a, for all a) and the assumed cumulative distribution function G_e. It is $sup_a|G_o(a) - G_e(a)|$.
- The **Lilliefors test** is for the assumption of normality with unspecified parameters, which makes it suitable to test general assumptions about normality. The test is based on the Kolmogorov–Smirnov statistic. The assumed cumulative distribution function is the normal distribution of the mean equal to the sample mean $\bar{Y}$ and variance equal to the sample variance (i.e. the unbiased estimator of the variance). The Lilliefors test is less conservative than the Kolmogorov–Smirnov test (Gibbons, 2011).
- The **Anderson–Darling test** is another modification of the Kolmogorov–Smirnov test that assigns more weight to the tails of the distribution. The test is more sensitive than the Kolmogorov–Smirnov test.
- The **Shapiro–Wilk test** statistic is $\mathbf{T}(\mathbf{Y}) = \frac{(\sum_k^n a_k Y_{(k)})^2}{\sum_i^n (Y_i - \bar{Y})^2}$, where the $Y_{(k)}$ are the ordered sample values from $k = 1 \ldots, n$ and the a_k are constants generated from the expected values and covariance matrix of the order statistics of a sample of size n from a normal distribution.

Razali and Wah (2011) compare the power of these four tests of normality via Monte Carlo simulation of sample data generated from alternative distributions that follow symmetric and asymmetric distributions. Results show that the Shapiro–Wilk test is the most powerful normality test, followed by the Anderson–Darling test, the Lilliefors test and the Kolmogorov–Smirnov test. Normality tests remain difficult to use for a couple of reasons. First, for small samples, normality tests usually have small power (Razali and Wah, 2011). Second, for large sample sizes, i.e. when the tests are supposed to have enough power, basic parametric tests (t-test or ANOVA) are quite robust to non-normality.

Focus 7.14
Testing for randomness: the run test

The run test offers a simple procedure to decide if the observed data are in random order. The observed sample $\mathbf{y} = (y_1, \ldots, y_n)$ corresponds to the order in which the data were obtained. The two hypotheses are simply:

$$\begin{cases} H_0: & \text{The data come in random order;} \\ H_1: & \text{The data do not come in random order.} \end{cases}$$

The test is based on the comparison between the number of runs of consecutive values and the median (or, more generally, a given threshold) of the observed data. This comparison yields a binary variable depending on whether a given value is over (n_u) or below (n_l) the median. A 'run' is defined as a group of successive identical values in a list, which will be evaluated using this binary variable. The alternative hypothesis is two-tailed as there is no particular reason to think that any departure from randomness will produce too many or two few runs. In other words, a two-tailed alternative hypothesis tests for both trend and cyclical effects. In a run test, the test statistic is the number of observed runs using this binary variable. The distribution of the run statistic has the probability distribution:

$$Pr[\mathbf{T}(\mathbf{Y}) = \mathbf{T}(\mathbf{y})] = \begin{cases} \dfrac{2\binom{n_u-1}{\mathbf{T}(\mathbf{y})/2-1}\binom{n_l-1}{\mathbf{T}(\mathbf{y})/2-1}}{\binom{n_u+n_l}{n_u}}, \text{ if } \mathbf{T}(\mathbf{y}) \text{ is even} \\ \dfrac{\binom{n_u-1}{(\mathbf{T}(\mathbf{y})-1)/2}\binom{n_l-1}{(\mathbf{T}(\mathbf{y})-3)/2}+\binom{n_u-1}{(\mathbf{T}(\mathbf{y})-3)/2}\binom{n_l-1}{(\mathbf{T}(\mathbf{y})-1)/2}}{\binom{n_u+n_l}{n_u}}, \text{ if } \mathbf{T}(\mathbf{y}) \text{ is odd} \end{cases}$$

The exact run test can be tedious to calculate. If n_u and n_l are both over 10, then the sample is considered to be large and the distribution of the test statistic $\mathbf{T}(\mathbf{Y})$ is approximately normal, $\mathbf{T}(\mathbf{Y}) \sim \mathcal{N}(\mu, \sigma^2)$ where:

$$\mu = \frac{2n_u n_l}{n_u + n_l} + 1 \text{ ; and } \sigma^2 = \frac{2n_u n_l(2n_u n_l - n_u - n_l)}{(n_u + n_l)^2(n_u + n_l - 1)}$$

Example: Consider the sample:

$$\mathbf{y} = (15, 16, 23, 12, 7, 14, 13, 13, 12, 9, 18, 5, 7, 9, 10, 12, 12, 11, 14, 15, 16, 13, 12, 9, 17)$$

and let h denote values above the median (which is 12.5) and l values below the median; we obtain a series with $\mathbf{T}(\mathbf{y}) = 9$ *runs:*

$$(\underbrace{h,h,h,}_{run}\ \underbrace{l,l,}_{run}\ \underbrace{h,h,h,}_{run}\ \underbrace{l,l,}_{run}\ \underbrace{h}_{run},\ \underbrace{l,l,l,l,l,l,l,}_{run}\ \underbrace{h,h,h,h,h,}_{run}\ \underbrace{l,l,}_{run}\ \underbrace{h,h}_{run})$$

The exact t-test p-value is 0.045 and the null hypothesis of randomness is rejected. Assuming that the test statistic is normally distributed, with $\mu = 14$ *and* $\sigma^2 = 6.24$*, produces the same p-value.*

The run test is the classic test of randomness. A one-tailed run test can be used to test for the presence of a trend in the data. A simple alternative to the run test is the sign test, which is based on the sign of the differences between two consecutive values in the sample. Other alternatives have been developed to deal with trends. These include the Mann–Kendall test, which is based on a Kendall's τ correlation test between time and the observed data, and Bartel's rank test that is based on the sum of squares of the differences in ranks of successive elements. Gibbons (2011) describes these alternatives at length.

The binomial test provides a statistical answer to this question. The exact binomial p-value for the test is $Pr[\mathbf{T}(\mathbf{Y}) \geq 19|n = 54, \theta = 0.5] = \sum_{i=19}^{54} \binom{54}{i} 0.5^i 0.5^{54-i} = 0.02$. *The decision rule leads to the rejection of the null hypothesis and the conclusion that the distribution is biased towards males.*

For large enough n, a normal approximation can be used. If the sample variance is measured by the hypothesised proportion under the null hypothesis, θ_0, the test statistic

$$\mathbf{T}(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0 - \frac{1}{2n}}{\sqrt{\theta_0(1-\theta_0)/n}}$$

is called the 'score-test statistic'. The quantity $-\frac{1}{2n}$ is used as a continuity correction. If the sample variance is, rather, measured by the estimator of the proportion, $\hat{\theta}$, the test statistic

$$\mathbf{T}(\mathbf{Y}) = \frac{\hat{\theta} - \theta_0 - \frac{1}{2n}}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}}$$

is called the 'Wald test statistic'. Both statistics are distributed normally. The score test performs better than the Wald test and is usually the preferred approximation to the binomial test.

Multinomial test

When there are more than two categories, the binomial test can be extended to a multinomial test. The null hypothesis is that all categories are equal to the specified values. The sample variables Y_i correspond to the categories. The test statistic is the number of times each category occurs in the sample and its sampling distribution is a multinomial distribution. As the calculation of multinomial probabilities can be tedious, this test is often approximated by a chi-squared test.

Example: Consider an experiment in which 200 participants choose between three alternatives: A, B and C. A total of 53 participants choose A, 112 choose B and the remaining 35 choose C. According to a theory, let's say the proportion of choices of alternative A should be 1/4*, those of B* 1/2 *and those of C* 1/4*. Answering the question 'Are the observed proportions* $(53/200, 112/200, 35/200)$ *significantly different from what would be expected?' amounts to the test:*

$$\begin{cases} H_0: & \text{The proportions are } \theta_A = \frac{1}{4}, \theta_B = \frac{1}{2}, \theta_C = \frac{1}{4}; \\ H_1: & \text{The proportions are different from } \theta_A = \frac{1}{4}, \theta_B = \frac{1}{2}, \theta_C = \frac{1}{4}. \end{cases}$$

The multinomial test allows us to decide between these hypotheses. The exact binomial p-value for the test is 0.04. For comparison purposes, a likelihood-ratio test produces a p-value of 0.037 and the standard chi-squared test a p-value of 0.047.

Chi-squared goodness-of-fit test

For large enough samples, a popular alternative to the binomial and multinomial tests is the chi-squared test. This test owes its popularity to the fact that it quickly approaches its asymptotic distribution. The chi-squared test is a goodness-of-fit test between the

observed distribution G_o and a theoretical (or expected) distribution G_e. The variables are categorical, and grouped into K mutually exclusive numerical classes $k = 1, \ldots, K$. The test hypothesis is

$$H_0 : G_o(Y_i) = G_e(Y_i), \forall i \text{ vs. } H_1 : G_o(Y_i) \neq G_e(Y_i) \text{ for some } i$$

The test statistic is

$$\mathbf{T}(\mathbf{Y}) = \sum_{k}^{K} \frac{(\bar{y}_k^o - \bar{y}_k^e)^2}{\bar{y}_k^e}$$

where $\bar{y}_k^o$ and $\bar{y}_k^e$ are the observed and expected frequencies for the kth class. Under the null hypothesis and for large samples, the test statistic $\mathbf{T}(\mathbf{Y})$ is distributed chi-squared with $K - 1$ degrees of freedom. The rejection region is such that $\mathbf{T}(\mathbf{Y}|H_0) > c$, with: $c = \chi_\alpha^2$.

With only two categories, the theoretical number of observations is simply $n \times \theta_0$, where θ_0 is the assumed proportion of the relevant event under the null hypothesis. As the chi-squared test is a goodness-of-fit test, it can be used to test more general distributions than the binomial and multinomial distributions. If the theoretical distribution contains m parameters to be estimated, the number of degrees of freedom has to be reduced accordingly to $K - 1 - m$. The large-sample approximation of the test statistic can be used as long as every expected frequency $\bar{y}_k^e$ is larger than 5.[18]

Example: Based on the example described above for the binomial test, the sample chi-squared statistic is $\mathbf{T}(\mathbf{y}) = \frac{(19-54\times0.5)^2}{54\times0.5} + \frac{(19-54\times0.5)^2}{54\times0.5} = 4.74$. *Based on a 5% type I error, the rejection threshold is* $c = 3.84$. *As the sample statistic is inside the rejection region, we can reject the null hypothesis. The p-value is* $Pr[\mathbf{T}(\mathbf{Y}) > \mathbf{T}(\mathbf{y})] = 0.029$.

7.3.4 Independent Sample Tests

In this subsection, we focus on hypothesis tests based on mutually independent random samples. These tests are useful for the comparison of treatments based on different samples. The samples are supposed to be randomly drawn independently of each other. Independence is assumed here not only within-sample but also between-sample. More formally, denote $\mathbf{y^1} = (y_1^1, \ldots, y_{n_1}^1)$ and $\mathbf{y^2} = (y_1^2, \ldots, y_{n_2}^2)$, the two samples of observations, drawn independently. Each y_i^k $(k = 1, 2)$ is assumed to result from a draw of a random variable Y_i^k. Elements of each sample are independent draws from the same population distribution: F_1 for Y_i^1 and F_2 for Y_i^2.

Interval and Normal Variables

Difference between two means

There are different ways to test the difference between the means θ_1 and θ_2 for approximatively normally distributed variables, depending on the assumptions made about the

[18] If the expected frequencies are less than 5, the usual procedure is to regroup adjacent groups until the expected frequency is greater than 5. Lower, less conservative thresholds for the minimum expected frequency, e.g. 2 to 5, can also be used.

variances in the populations. If the two population variances are unknown, but assumed to be the same, the test is a t-test (with pooled variance). If the population variances are unknown, but assumed to be different, the test is a Welch test, with each variance estimated separately. If the variances are rather assumed to be known, the appropriate test is a z-test. In all three cases, the hypotheses are:

$$H_0 : \theta_1 = \theta_2 \text{ versus } H_1 : \theta_1 \neq \theta_2$$

For the t-test, the estimator of the difference between the means is $\bar{Y}^1 - \bar{Y}^2$ and the test statistic writes:

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \frac{(\bar{Y^1} - \bar{Y^2}) - (\theta_1 - \theta_2)}{S_p\sqrt{(1/n_1) + (1/n_2)}}, \text{ with } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

The sampling distribution under the null hypothesis,

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}|H_0) = \frac{\bar{Y^1} - \bar{Y^2}}{S_p\sqrt{(1/n_1) + (1/n_2)}}$$

has a t-distribution with $(n_1 + n_1 - 2)$ degrees of freedom. The rejection region is such that $c_l = -t_{\frac{\alpha}{2}}$ and $c_u = t_{\frac{\alpha}{2}}$, where t is the t-score with $n_1 + n_2 - 2$ degrees of freedom. If the variance σ^2 is known, the test is a z-test with common variance. The critical values of the test statistic are replaced by the corresponding z-scores, and the estimator of the pooled variance S_p^2 is replaced by σ^2.

In these tests, the alternative hypothesis H_1 can be interpreted in terms of location and stochastic dominance. The alternative hypothesis assumes that the populations are of the same shape, but with a different measure of central tendency. This corresponds to a test of location and allows stochastic dominance to be tested. For example, a one-sided test $H_1 = \theta_1 > \theta_2$ tests if the Y_i^1s are stochastically larger than the Y_i^2s.

If the variances of the two population distributions are not supposed to be equal, the difference between two means is tested via a Welch's two-sample t-test. In this case, the test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \frac{(\bar{Y^1} - \bar{Y^2}) - (\theta_1 - \theta_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Under the null hypothesis, this test statistic has a t-distribution with degrees of freedom equal to

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Bigg/ \left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)$$

If the variances σ_X^2 and σ_Y^2 are known, the test of the difference between two means is a z-test. The critical values of the test statistic are calculated accordingly and the estimated variances S_1 and S_2 are replaced by the known variances σ_1^2 and σ_2^2. The Welch test has less power than the t-test as fewer data are used to estimate the variances.

Example: Consider that two experimental sessions from second-price auction experiment have been run in two different universities, and that 54 people participated in the first session while there were 65 participants in the second session. You suppose that both population distributions are approximatively normal. A two-sample test allows us to test whether the payoffs were drawn from populations with the same mean. Suppose the sample means are $\bar{y}^1 = 8.89$ *and* $\bar{y}^2 = 8.24$*, and the sample standard deviations are* $s_1 = 2.76$ *and* $s_2 = 3.38$*. The test statistic is* 1.135 *for the two-sample t-test and* 1.156 *for the two-sample Welch test. In both cases, the critical value at* $\alpha = 0.05$ *is* $t_{0.025} = 1.98$ *for a t-score with 117 degrees of freedom. According to the decision rule, hypothesis* H_0 *is not rejected and the mean observed payoffs are the same between sessions.*

The ratio between two variances: the F-test

The F-test is the dispersion version of the two-sample comparison of means. The F-test assesses whether one population varies more than another. The statistic is the ratio of the sample variances, $\frac{S_1^2}{S_2^2}$. The sampling distribution under the null hypothesis is a Fisher distribution F_{n_1-1,n_2-1}. The rejection region is such that:

$$c_l < F_{n_1-1,n_2-1,1-\frac{\alpha}{2}} \text{ or } c_u = F_{n_1-1,n_2-1,\frac{\alpha}{2}}$$

The F-test is not robust to departures from normality. If non-normality is suspected, non-parametric tests are preferred as rejection of the null hypothesis may well correspond to a rejection of normality. One advantage of the F-test is that it does not require any assumption about the mean of the normal populations, which is not the case for the other parametric tests, such as the t-test. If the population means are assumed to be equal, the alternative hypothesis in the F-test can be interpreted in terms of scale and second-order stochastic dominance. Under this assumption, the hypothesis H_1 assumes that the populations are of the same shape, but with a different measure of dispersion. This tests for second-order stochastic dominance, and in particular for mean-preserving spreads. For example, a one-sided test $H_1 = \theta_1^2 > \theta_2^2$ tests whether the population distribution of the Y_i^1s is a mean-preserving spread of that of the Y_i^2s.

The F-test is based on an estimator of the ratio of the sample variances. A Wald test can be constructed to test the differences between the variances instead of testing their ratio. In this case, the Wald test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \frac{(S_1^2 - S_2^2)^2}{\frac{2S_1^4}{n_1+1} + \frac{2S_2^4}{n_2+1}}$$

and is (asymptotically) chi-squared distributed with one degree of freedom. The F-test and the Wald test for the difference between variances are equivalent.

Difference between K means: one-way ANOVA

The comparison of the means of two normally distributed populations performed by the t-test can be extended to the comparison of the means of K normally distributed populations with the same variances. Pairwise comparisons cannot be used with $K >$

Table 7.6 The ANOVA decomposition

Source of variation	d.f.	Sum of squares	Mean squares	F
Group	$K-1$	$SS_{between}$	$S^2_{between}$	$\frac{S^2_{between}}{S^2_{within}}$
Error	$n-K$	SS_{within}	S^2_{within}	
Total	$n-1$	SS_{total}		

2 treatments as the comparisons are not independent of each other. The logic of the analysis of variance (ANOVA) is to test a single hypothesis of equality between the K means, for independent draws from a normal distribution with common unknown variance σ^2. The hypothesis under investigation is

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_K \text{ vs. } H_1 : \text{ the means are different for some } i, j, i \neq j.$$

If H_0 is rejected, then additional inference is required to identify the source of the rejection. The ANOVA is based on a comparison of the means and variances in the different samples. If the variation in the sample means is large relative to the variance within samples, then the observations suggest the rejection of the null hypothesis. Formally, the total dispersion (the sum of the squares of the gaps between the observed values and overall mean) SS_{total} is decomposed into two components:

$$SS_{total} = SS_{within} + SS_{between}$$

The first, SS_{within}, is the sum of squares of the gaps between the observed values Y_i^k and their sample group means ($\bar{Y}_k = \frac{\sum_i Y_i}{n_k}$). The second, $SS_{between}$, is the weighted sum of squares of the gaps between the sample group means and the overall mean, with the weights being the number of observations in each sample n_k. The ratio of $SS_{between}$ to the total sum of squares SS_{total} is the coefficient of determination (the 'R-squared' value). The estimate of the sample variance within each group is SS_{within} divided by $n-K$, with $n = \sum_k n_k$. The estimate of the sample variance within each group is $SS_{between}$ divided by $K-1$. In the analysis of variance, these estimates of the sample variance are also called the mean sum of squares, defined as:

$$S^2_{within} = \frac{\sum_k \sum_i (Y_i^k - \bar{Y}_k)^2}{n-K}, \; S^2_{between} = \frac{\sum_k n_k (\bar{Y}_k - \bar{\bar{Y}})^2}{K-1}$$

The test statistic is the ratio $S^2_{between}/S^2_{within}$, and its sampling distribution under the null hypothesis is a Fisher distribution $F_{K-1,n-K}$. The rejection region is $S^2_{between}/S^2_{within} > c$, with c such that $c = F_{K-1,n-K,\alpha}$. If the null hypothesis is not rejected, then $\bar{\bar{Y}}$ is an estimate of the mean and s^2_{within} is a sample estimate of the (hypothesised) common variance. The sample variance $S^2_{between}$ is sometimes referred to as the sampling error, whereas the sample variance S^2_{within} is sometimes called the mean squared error, the experimental error or the pooled estimate of the variance. Table 7.6 shows the decomposition of the components of the ANOVA test.

If the null hypothesis is rejected, the question arises as to which group of means is behind the rejection of the null hypothesis. Multiple-comparison, or *post hoc*, procedures systematically compare pairs of means to establish the cause of the rejection of the null hypothesis. One intuitive procedure is the least-significant difference. This procedure is based on the calculation of the smallest significant difference between the means of two groups k and k' with a t-test. Any difference between two means $\bar{Y}_k - \bar{Y}_{k'}$ that is greater than the smallest significant difference is taken to be significant. The least-significant difference is

$$LSD_{k,k'} = t_{\alpha/2}\sqrt{S^2_{within}\left(\frac{1}{n_k} + \frac{1}{n_{k'}}\right)}$$

where t is the t-score with $n - K$ degrees of freedom. One difficulty with this method is that it does not correct for multiple comparisons and inflates the type I error. Two popular ways of addressing this issue are the Tukey and Scheffé methods. The Tukey method tests all possible pairs of differences to find significant differences, correcting for multiple comparisons. This amounts to replacing the t statistic in the LSD formula by the (standardised) difference between the largest and smallest of the K means, $\bar{Y}_{max}$ and $\bar{Y}_{min}$ under H_0. This is equal to:

$$\mathcal{Q} = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{\sqrt{S^2_p\left(\frac{1}{n_{k_{max}}} + \frac{1}{n_{k_{min}}}\right)}}$$

where S_p is the estimate of the pooled variance of the means $\bar{Y}_{max}$, and $\bar{Y}_{min}$ and $n_{k_{max}}$, $n_{k_{min}}$ are the corresponding number of observation in each group. The distribution of the variable $\mathcal{Q}$ is called the studentised range distribution with K groups, $n - K$ degrees of freedom and significance of α. The corresponding tabulated value $\mathcal{Q}_{K,n-K,\alpha}/\sqrt{2}$ serves as a corrected $t_{\alpha/2}$ value to test for significant differences between pairs of means. This method is also called Tukey's honestly significant difference (HSD).

The Scheffé method tests all possible comparisons of groups of means. With $K > 2$ treatments, the rejection of the null hypothesis often does not suffice to draw a firm conclusion regarding the difference between treatments. Consider an experiment with one control group (θ^1) and two treatment groups (θ^2 and θ^3). The rejection of the equality of means between treatments leaves some questions unanswered, for instance whether H_0 was rejected because the control group is different from the average of the treatment groups ($\theta^1 \neq (\theta^2 + \theta^3)/2$), or because the two treatment groups are different ($\theta_2 \neq \theta_3$). These comparisons between groups of parameters are called *contrasts* in ANOVA. When the comparison is between two means (e.g. $H_0 : \theta_2 \neq \theta_3$), H_0 is called a simple contrast hypothesis; when it involves more than two means (e.g. $H_0 : \theta^1 = (\theta^2 + \theta^3)/2$), it is called a complex contrast hypothesis. The test statistics here are called contrast statistics. As the contrast hypothesis is based on a linear combination of population means, the sampling distributions of the contrast statistics are a linear combination of normal distributions.

The ANOVA can be represented as a special case of a linear model $Y_i^k = \theta_k + \epsilon_{ik}$ in which the dependent variables are the Y_i^k and the independent variable(s) is (are) the

Focus 7.15
Two-way and multi-way ANOVA

ANOVA can be generalised to experiments where two (or more) categorical variables are used to define between-subject treatments. For example, suppose the second-price auction experiment has been run in three different universities, where in each session subjects are assigned at random to incentivised or hypothetical treatments. In this case, each subject faces a particular combination of localisation and incentives, which are the explanatory variables. In ANOVA, the explanatory variables are often called 'factors' and their values are called 'levels'. In this example, there are two factors (factor A: universities, factor B: incentives), with three levels for the first and two for the second. If there are the same number of subjects for each possible combination of the explanatory variables, the design is 'balanced'. In our example, the design is a between-subject 3×2 design. With more than two explanatory variables, the ANOVA is 'multi-way'. Two-way ANOVA tests the effect of each factor with or without interactions. Without interactions, the model is called 'additive': in the example, this results, for instance, from the assumption that the bid in the incentivised treatment does not depend on the university in which the session is run. The assumptions for the multi-way analysis of variance are very close to those for standard one-way ANOVA. Each observation is assumed to be independently drawn from a normal distribution with common unknown variance σ^2 and the explanatory variables (factors) are categorical. The test statistic is

$$\frac{S^2_{between}}{S^2_{within}}$$

for each explanatory variable and each interaction between the explanatory variables (if any). In a two-way ANOVA with interaction effects, there are three hypothesis tests: the main effect of factor A (equality of means); the main effect of factor B (equality of means); and the interaction between factors A and B. Multi-way ANOVA is often run with all interactions included: if the interactions are insignificant, then an additive model is rerun. A two-way ANOVA can be represented as a linear model:

$$Y_i^{k,j} = \underbrace{\theta}_{\text{overall mean}} + \underbrace{a_k}_{\text{effect of factor A}} + \underbrace{b_j}_{\text{effect of factor B}} + \underbrace{(ab)_{kj}}_{\text{interaction effect}} + \epsilon_{ikj}$$

where ϵ_{ikj} is a $\mathcal{N}(0, \sigma^2)$ normally distributed error.

treatment, or grouping, variable. For example, an ANOVA is equivalent to the linear regression

$$Y_i^k = a_1 + a_2 Z_2 + \ldots + a_n Z_n + \epsilon_{ik}$$

where Y_i^k are the observations, ϵ_{ik} is a $\mathcal{N}(0, \sigma^2)$ normally distributed error and $Z_k, k = 2, \ldots, K$ are dummies for the subject being in treatment k. Here, a_1 is an estimate of the mean of the reference treatment $k = 1$ and $a_k, k = 2, \ldots, K$ are the estimates of the differences in means between treatment k and the reference treatment. Focus 7.15 generalises the ANOVA to situations where categorical variables define between-subject treatments.

Difference between K variances: Bartlett test

The Bartlett test is a modification of the likelihood-ratio test. It is used to test for the homogeneity of variance (homoskedasticity) between K populations. Denoting θ_k the variance for population k, the hypothesis under investigation is

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_K \text{ vs. } H_1 : \theta_i = \theta_j \text{ are different for some } i, j, i \neq j$$

The test statistic $\mathbf{T}(\mathbf{Y^1}, \ldots, \mathbf{Y^K})$ is:

$$\frac{(\sum_k n_k - K)\log(S^2_{within}) - \sum_k (n_k - 1)\log(S^2_k)}{1 + \frac{1}{3(k-1)\left(\sum_k \frac{1}{n_k - 1} - \frac{1}{\sum_k n_k - K}\right)}}$$

The sampling distribution under the null hypothesis is chi-squared with $K - 1$ degrees of freedom. The rejection region is an upper rejection region with $c = \chi^2_{K-1,\alpha}$.

The Bartlett test is not robust to departures from normality. Under non-normality, alternative tests (the parametric Levene test or the Brown–Forsythe test) are preferred, as rejection of the null hypothesis may result from the departure from normality.

Interval or Ordinal Variables

Difference between distributions: the Mann–Whitney and Kolmogorov–Smirnov tests

The Mann–Whitney test, or U-test, is particularly useful for testing for the difference between two distributions. When two populations are assumed to be identical except for their location, the Mann–Whitney test can be used to compare their means or medians. The small number of assumptions required for this test explains its popularity. One important characteristic of the test is that small sample sizes produce an accurate normal approximation. The test statistic (generally denoted U) is based on two samples Y^1_i, Y^2_i and counts the number of times a Y^2_i precedes a Y^1_i in the ordered combination of the two samples. The test assumes that the random variables Y^1_i, Y^2_i are independent *i.i.d.* draws from populations with location parameters θ_1 and θ_2. The hypothesis under investigation is

$$H_0 : G_1(a) = G_2(a), \forall a \text{ vs. } H_1 : G_1(a) \neq G_2(a) \text{ for some } a$$

The test statistic is $\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \sum_i \sum_j Z_{ij}$, where Z_{ij} is a binary variable taking a value of 1 whenever $Y^2_j < Y^1_i$ and 0 otherwise. Under the null hypothesis the sampling distribution is based on that of the Bernoulli variables Z_{ij}. The rejection region is

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}|H_0) < c \text{ or } \mathbf{T}'(\mathbf{Y^1}, \mathbf{Y^2}|H_0) < c$$

with $\mathbf{T}'$ defined over the binary variable taking a value of 1 whenever $(Y^2_j > Y^1_i)$ and 0 otherwise.

In practice, in the case of ties, Z_{ij} is assigned a value of 0.5. For large samples, the test statistic can be calculated by creating a pooled sample with observations $\mathbf{Y^1}$ and $\mathbf{Y^2}$ and assigning ranks to the observations in this grand sample. The test statistic for each sample is equal to the sum of the ranks for one sample, minus the sum of all possible ranks in that sample. For example, $\mathbf{T}_1(\mathbf{Y^1}, \mathbf{Y^2}) = \sum_i^{n_1} r_i(Y^1_i) - \frac{n_1(n_1+1)}{2}$ or

$\mathbf{T}_2(\mathbf{Y^1}, \mathbf{Y^2}) = \sum_i^{n_2} r_i(Y_i^2) - \frac{n_2(n_2+1)}{2}$. The test statistic under H_0 is the minimum between $\mathbf{T}_1(\mathbf{Y^1}, \mathbf{Y^2})$ and $\mathbf{T}_2(\mathbf{Y^1}, \mathbf{Y^2})$. This representation of the test statistic shows that the Mann–Whitney test is equivalent to a Wilcoxon signed-rank test. If the two groups are very different, then $\mathbf{T}_1$ (for example), will be equal to $\frac{n_1(n_1+1)}{2}$, $\mathbf{T}_2$ will be 0 and the test statistic will be 0. If the two groups are not very different, $\mathbf{T}_1 + \mathbf{T}_2$ will be close to $n_1 n_2$ and the test statistic will be close to $n_1 n_2/2$. For sample sizes $n_1 > 8$ and $n_2 > 8$, a normal approximation for the test statistic under H_0 is

$$\mathcal{N}\left(\frac{n_1 n_2}{2}, \sqrt{\frac{(n_1 n_2)(n_1 + n_2 + 1)}{12}}\right)$$

and the critical values are the usual z-scores. The one-sample Kolmogorov–Smirnov test can be adapted to test for identical distributions in a two-sample problem. The two samples are reordered to form an increasing sequence of values $Y^1_{(1)}, \ldots, Y^1_{(n_1)}$ and $Y^2_{(1)}, \ldots, Y^2_{(n_2)}$, and the empirical cumulative distributions G^1_o and G^2_o are constructed using these sequences. The test is based on the maximum absolute difference between these two empirical distributions. The hypotheses are identical to those for the Mann–Whitney test. The test statistic is $\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = max_a|G^1_o(a) - G^2_o(a)|$ (see Gibbons, 2011, for details about the sampling distribution and the rejection region).

Difference in the scale parameter: the Siegel–Tukey and Ansari–Bradley tests

The Siegel–Tukey rank-dispersion test is an important alternative to the F-test. The null hypothesis assumes that both samples come from identical populations, while the alternative hypothesis assumes that the two samples come from different populations where only variability (i.e. scale) differs. The Siegel–Tukey test is based on the Wilcoxon test. Two samples are merged in a grand sample, and ordered from the lowest to the largest. Ranks are then assigned to this ordered grand sample. Rank 1 is assigned to the smallest observation, rank 2 to the largest, rank 3 to the next-largest observation, rank 4 to the next-smallest observation, and so on. A rank-sum test is applied to the difference between the two populations, based on their ranks. If one group is more dispersed than the other, it will have more of the lower ranks that are assigned to the more extreme values. The test statistics are the Mann–Whitney test statistics for each group, minus $n(n+1)/2$, where n is the size of the grand sample. The Ansari–Bradley test is based on a slightly different test statistic, where the rank of 1 is assigned to both the smallest and the largest observations, 2 to the second-smallest and second-largest observations, and so on.

Difference between K samples: the Kruskal–Wallis test

The Kruskal–Wallis test is a generalisation of the Mann–Whitney test for the comparison of K independent samples. The test statistic is based on the ranks of each observation Y_i^k in the rank-ordered sequence of all observations, and is the weighted sum of squares of the gap between the observed and expected rank sums. The null hypothesis assumes that the K samples are drawn from the same common population. The test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \ldots, \mathbf{Y^K}) = \frac{12}{\sum_k n_k(\sum_k n_k + 1)} \sum_{k=1}^{K} \frac{\sum_i r(Y_i^k)^2}{n_k} - 3(\sum_k n_k + 1)$$

Table 7.7 A 2×2 table for independent samples

	Successes	Failures	Total
Population 1	a_1	b_1	$a_1 + b_1$
Population 2	a_2	b_2	$a_2 + b_2$
Total	$a_1 + a_2$	$b_1 + b_2$	n

The sampling distribution under the null hypothesis for large samples ($n_k \geq 5, \forall k$ and $K \geq 4$) is chi-squared with $K - 1$ degrees freedom. The rejection region is an upper rejection region with $c = \chi^2_{K-1,\alpha}$.

If the null hypothesis is rejected, pairwise comparisons between samples can be carried out based on the average ranks in each sample. The treatments k and k' are significantly different if:

$$\left| \frac{\sum_i r(Y_i^k)}{n_k} - \frac{\sum_i r(Y_i^{k'})}{n_{k'}} \right| \geq z_{1-\alpha} \sqrt{\frac{n(n+1)}{12} \left(\frac{1}{n_k} + \frac{1}{n_{k'}} \right)}$$

with $n = \sum_k n_k$. If the alternative hypothesis can be specified as an increasing order of location parameters (means or medians), $H_1 : \theta_1 \leq \theta_2 \leq \ldots \leq \theta_K$, the Jonckheere–Terpstra test for pairwise comparisons is used. Gibbons (2011) presents a number of procedures to test the partial ordering of the $K - 1$ distributions.

Another way of carrying out multiple comparisons is the Dunn test. This compares two groups k and k' based on the absolute value of the difference between their mean, divided by the quantity $\sqrt{[n(n+1)/12][(1/n_k) + (1/n_{k'})]}$. The test statistic is normally distributed under the null hypothesis, and the standard z-scores apply for the critical values of the rejection regions.

Difference between K variances: the non-parametric Levene test

The Levene test can be used to test the equality of K variances when the sample is not assumed to be normally distributed. In this case, the test statistic is based on the rank of each observation Y_i^k in the rank-ordered sequence of all observations. The general idea of the test is to carry out an ANOVA on the absolute difference between the rank of each observation and the mean of these ranks. If the median (or a trimmed mean) is used instead of the mean, the test statistic is a non-parametric Brown–Forsythe test, which is more accurate when the underlying distributions are not symmetric.

Categorical Variables

Fisher exact test

When two samples are independent draws from Bernoulli distributions with probabilities of success θ_1 and θ_2, the data can be represented in a 2×2 table. An example is shown in Table 7.7, which lists the success and failure counts for each sample. The quantities $a_1 = \sum_i^{n_1} Y_i^1$ and $a_2 = \sum_i^{n_2} Y_i^2$ denote the number of successes in each population and the quantities b_1 and b_2 the corresponding number of failures. The marginal probability of success is $(a_1 + a_2)/n$ and the marginal probability of failure $(b_1 + b_2)/n$.

The logic of the Fisher exact test is to evaluate the probability of observing the 2×2 table from the sample among all possible 2×2 tables that could have resulted with the same marginals (i.e. with the row and column totals fixed at their observed levels). Under the null hypothesis, the conditional distribution of a given the marginal totals is hypergeometric. The exact hypergeometric probability of observing the data in Table 7.7 is

$$\frac{\binom{a_1+b_1}{a_1}\binom{a_2+b_2}{a_2}}{\binom{n}{a_1+a_2}}$$

The Fisher exact test is particularly useful for small frequencies, where the chi-squared approximation might not hold. Moreover, and contrary to the Fisher exact test, the chi-squared test can only be used for a two-sided alternative hypothesis.

A z-test can be used to test the null hypothesis of identical proportions θ_1 and θ_2 if the sample sizes are large enough to justify a normal approximation. The test statistic under H_0 is normally distributed. It is defined as the ratio of the difference in sample proportions $\bar{y}^1 - \bar{y}^2$ to the estimator of the common proportion $\hat{\theta_0} = \frac{\sum_i^{n_1} Y_i^1 + \sum_i^{n_2} Y_i^2}{n_1+n_2}$.

Chi-squared test

When K samples are independently drawn from Bernoulli distributions with parameters θ_k, a chi-squared test can be used to test the null hypothesis of identical proportions. The alternative hypothesis is that at least two proportions differ $\theta_k \neq \theta_{k'}$ for some (k, k'). The test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \ldots, \mathbf{Y^K}) = \sum_k^K \frac{(\bar{y}^k - n_k\theta_k)^2}{n_k\theta_k(1-\theta_k)}$$

where $\bar{y}^k = \sum_i y_i^k$ is the number of successes in sample k. Under the null hypothesis the estimator of the common proportion is $\hat{\theta_0} = \sum_k \bar{Y}^k/n$. Under H_0, the test statistic is distributed chi-squared with $K-1$ degrees of freedom, and the rejection region is an upper rejection region with $c > \chi^2_{K-1,\alpha}$. When applied to Table 7.7, the test statistic from the chi-squared test is

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}|H_0) = \frac{(n-1)(a_1b_2 - b_1a_2)^2}{(a_1+b_1)(a_1+a_2)(b_1+b_2)(a_2+b_2)}$$

and is distributed chi-squared with one degree of freedom.

7.3.5 Paired Samples and Repeated Measure Tests

When the data are repeated measures over the same observation unit, or when observation units are matched together, the observations are no longer independent – an important validity assumption for the tests presented in the previous subsection. This is the case when the same subjects face different treatments or when they face repetitions of the same treatment a number of times during the experiment. With paired data, inferences are drawn from the gaps between pairs of observations, and the two-sample tests are more similar to one-sample tests on the (matched) population differences.

With repeated measures, statistical tests have to take into account the within-subject correlation in the measures.

Interval and Normal Variables

Difference between two means: the paired t-test

The paired t-test evaluates the difference between treatment effects in matched pairs. Under the null hypothesis the mean of the paired differences θ is 0. The assumption of normality does not refer to the population distribution of the variables, but rather to their differences. The sample is assumed to consist of draws of paired random variables Y_i^1 and Y_i^2, with differences $\Delta_i = Y_i^1 - Y_i^2$ that are *i.i.d.* normally distributed, with unknown variance σ^2. The null hypothesis is $H_0 : \theta = 0$. The estimator of the mean of the difference is $\bar{\Delta} = \bar{Y^1} - \bar{Y^2}$, and the test statistic

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \frac{(\bar{\Delta}) - \theta}{S_\Delta} \text{ with } S_\Delta^2 = \frac{\sum_i (\Delta_i - \bar{\Delta})^2}{n-1}$$

is t-distributed with $n-1$ degrees of freedom under the null hypothesis. The rejection region is such that $c_l < -t_{\frac{\alpha}{2}}$ or $c_u > t_{\frac{\alpha}{2}}$, where t is the t-score with $n-1$ degrees of freedom.

Difference between K means: the repeated-measure ANOVA

The repeated-measure ANOVA is a generalisation of the paired t-test when the same individual has K different treatments. The repeated-measure ANOVA takes into account that the errors in the multiple measurements of each participant are correlated. The paired t-test estimator is based on the difference between the sample means. As a result, no hypothesis is made about the correlation between two treatments. However, with $K > 2$ correlation does become important. Here, the repeated Y_i^k are assumed to be normally distributed for each $k = 1, \ldots, K$ and *i.i.d.* between subjects. Due to the repeated treatments, the normal distribution is multivariate, so that the vector of responses $\left(Y_i^1, \ldots, Y_i^K\right)$ is drawn independently from the same population distribution:

$$\left(Y_i^1, \ldots, Y_i^K\right) \sim \mathcal{N}\left[\left(\theta^1, \ldots, \theta^K\right), \Omega\right]$$

where Ω is the within-subject variance–covariance matrix. If we suppose that the variances are equal across the K treatments and that there is no interaction between treatments (0 covariance), then the measurements are assumed to be independent. If all variances are equal and all covariances are equal then compound symmetry is assumed. In this case, the interactions are the same throughout the experiment and the variance of the difference between treatments is constant. A less-restrictive form of compound symmetry is sphericity. This holds when variances of the differences between treatments are equal. If sphericity does not hold, then the variance–covariance figures interact with the difference between means, and have to be estimated. This additional estimation reduces the power of the test. The interaction is taken into account through the 'correction' of the repeated-measure ANOVA. The Huynh–Feldt correction is standard and is provided by most statistical packages.

Difference between two variances

The equality of two variances in paired samples can be evaluated with a test statistic based on the Pearson correlation between the variables $Y_i^1 - Y_i^2$ and $Y_i^1 + Y_i^2$. The test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

where ρ is the sample estimate of the Pearson correlation between $Y_i^1 - Y_i^2$ and $Y_i^1 + Y_i^2$. The sampling distribution under the null hypothesis is t-distributed with $n-2$ degrees of freedom. The rejection region is such that $c_l < -t_{\frac{\alpha}{2}}$ or $c_u > t_{\frac{\alpha}{2}}$, where t is the t-score with $n-2$ degrees of freedom.

Interval or Ordinal Variables

Difference between two medians: the Wilcoxon matched-pairs test and sign test

The Wilcoxon matched-pairs signed-rank test establishes whether the difference between pairs of observations are symmetrically distributed about a median of 0. The test assumes that the differences $Y_i^1 - Y_i^2$ are symmetric around the median θ. The null hypothesis is $H_0 : \theta = 0$ and the alternative hypothesis is $H_1 : \theta \neq 0$. The test statistic is the expected value of the sum of the ranks for positive ($\mathbf{T}^+$) and negative ($\mathbf{T}^-$) differences $\Delta_i = Y_i^1 - Y_i^2$ and the logic of the one-sample Wilcoxon test is applied to these differences.

Both the paired t-test and the Wilcoxon test assume that the pairs of observations are drawn from the same common population. The sign test does not require this hypothesis. In the sign test for matched pairs, the null hypothesis assumes that the average differences of paired observations is 0. The test statistic is based on the counts of positive and negative differences between Y_i^1 and Y_i^2. With the sign test, H_0 is rejected if the number of differences is too large or too small.

Difference between k medians: the Friedman test

The Friedman test supplies a non-parametric analysis of the variance in repeated-measure designs. The null hypothesis is that the treatment effects are equal, $H_0 : \theta_1 = \ldots = \theta_K$, and the alternative hypothesis is that at least two medians are different. The test assigns, for each subject (each block), a rank in the $1, \ldots, K$ sequence for each observation and calculates the sums R_k of the ranks for each treatment. Under the null hypothesis, each rank within each block is equally likely. It is assumed that the blocks are mutually independent and that there are no ties within each block. The test statistic is the sum of squares of the gaps between these sums of ranks and their mean:

$$\mathbf{T}(\mathbf{Y^1}, \ldots, \mathbf{Y^K}) = \frac{12}{nK(K+1)} \sum_{k=1}^{K} R_k^2 - 3n(K+1)$$

and is distributed chi-squared with $K-1$ degrees of freedom. The test statistic needs to be corrected when there are ties. The rejection region is an upper rejection region with $c = \chi^2_{K-1,\alpha}$. If H_0 is rejected, a Dunn test can be used to carry out multiple comparisons.

Table 7.8 A 2 × 2 table for paired samples

Treatment 1	Treatment 2 Success	Failure	Total
Success	a_{ss}	a_{sf}	$a_{ss} + a_{sf}$
Failure	a_{fs}	a_{ff}	$a_{fs} + a_{ff}$
Total	$a_{ss} + a_{fs}$	$a_{sf} + a_{ff}$	n

Categorical Variables

Difference between two proportions: the McNemar test

The McNemar test is used to test the impact of a treatment when the variables are categorical, with matched pairs of subjects. A typical example is observations on successes and failures before and after some treatment. Table 7.8 shows the counts of successes and failures for two repeated treatments over the same sample.

Under the null hypothesis the two treatments are similarly effective: then the two marginal probabilities of success are identical, as are the marginal probabilities of failure. The test then amounts to testing the similarity of the failure proportions between treatments 1 and 2. The test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \mathbf{Y^2}) = \frac{(a_{sf} - a_{fs})^2}{a_{sf} + a_{fs}}$$

Under the null hypothesis, with $a_{sf} + a_{fs} > 25$, the sample distribution is distributed chi-squared with one degree of freedom. If the cell sizes are too small for the chi-squared approximation, a binomial test can be used instead. One possible issue is that both the McNemar and binomial tests can be conservative for small samples. The remedy is to apply a downward correction to the p-value based on the binomial probability of observing a_{sf}. The test after correction is called a McNemar mid-P test.

Difference between K populations: the Cochran Q test

The Cochran Q test is a conservative test used to evaluate the difference between K treatments. Under the null hypothesis the K treatments are similarly effective; the alternative hypothesis is that at least two treatments differ. The test statistic is

$$\mathbf{T}(\mathbf{Y^1}, \ldots, \mathbf{Y^K}) = K(K-1)\frac{\sum_k \left(\sum_i Y_i^k - \frac{n}{K}\right)^2}{\sum_i \left[\sum_k Y_i^k (K - \sum_k Y_i^k)\right]}$$

Under the null hypothesis, and with a sufficiently large sample size, the test statistic is distributed chi-squared with $K - 1$ degrees of freedom. The rejection region is an upper rejection region with $c = \chi^2_{K-1,\alpha}$.

7.4 *Case Study*: Eliciting Preferences under Risk

The measurement of risk preferences has received a great deal of attention in both the experimental-economics and decision-analysis literatures. Both have proposed methods

and experimental designs to assess individual risk preferences. In experimental economics, the measurement of risk preferences is traditionally based on a series of discrete choices between uncertain risky prospects. Other approaches are based on allocation or investment decisions. In this case study we provide an overview of the measurement of risk preferences together with their estimation.

7.4.1 Expected-Utility Theory

The standard normative theory of decision under risk and uncertainty, expected-utility theory (von Neumann and Morgenstern, 1944; Savage, 1954), provides a set of consistent tools for the measurement of risk attitudes. The simplicity of the theory (expected utility corresponds to a weighted average) has made it by far the most prominent model of decision under risk and uncertainty in economics. Under expected utility, a decision-maker's attitude towards risk can be evaluated via a small number of equivalences between lotteries.

In what follows we consider a decision-maker who chooses between two lotteries. As most techniques only use lotteries with at most two distinct outcomes, this is what we present. Formally, let $x_1{}_p x_2$ denote a lottery producing outcome x_1 with probability p and x_2 with probability $1-p$. The individual has preferences over lotteries, and we use the conventional notation $\succ$, $\succsim$, and $\sim$ to represent the relations of strict preference, weak preference and indifference. The outcomes are real numbers, and typically represent money: larger outcome values are always preferred. If $x_1 = x_2$ the lottery is said to be riskless or certain, otherwise it is said to be risky. For the sake of clarity, we denote the riskless lottery x_c (hence $x_1 = x_2 = x_c$).

Expected utility under risk holds if the decision-maker's choices conform to three main behavioural axioms (Jensen, 1967; Fishburn, 1989). The first is a standard weak-order axiom, with $\succsim$ being complete and transitive. The second axiom is technical: continuity (see Gilboa, 2009, for a discussion of the behavioural content). Last, the main and by far the most important axiom is independence. The von Neumann and Morgenstern (1944) theorem shows that the preference relation satisfies these three axioms if and only if it can be represented by an expected-utility function.

Under expected utility, the decision-maker evaluates each lottery separately and chooses the lottery with the highest value. The valuation of a lottery is given by

$$U(x_1{}_p x_2) = pu(x_1) + (1-p)u(x_2) \tag{7.1}$$

where u is the utility function and U is the decision criterion, also called the von Neumann and Morgenstern (vNM) functional. The utility function u is a function from $\mathbb{R}$ to $\mathbb{R}$ and is unique up to an increasing affine transformation, i.e. up to intercept and slope. In (7.1), p and $1-p$ are the decision weights attached to the subjective values of outcomes x_1 and x_2. The decision criterion U has two main properties. The first is additivity: under expected utility, the utilities of the consequences u are combined additively. The second is linearity: the probabilities associated with the different outcomes enter the function linearly. These two properties together have important implications: empirically, they allow for the simple elicitation of preferences under risk. Under expected

Table 7.9 Elicitation methods

Standard-gamble methods	
1. Preference comparison	$x_{1_p}x_2$ versus x_c
2. Probability equivalence	$x_{1_{\underline{p}}}x_2 \sim x_c$
3. Value equivalence	$\underline{x_{1_p}}x_2 \sim x_c$
4. Certainty equivalence	$x_{1_p}x_2 \sim \underline{x_c}$
Paired-gamble methods	
5. Preference comparison	$x_{1_p}x_2$ versus $x'_{1_p}x'_2$
6. Probability equivalence	$x_{1_{\underline{p}}}x_2 \sim x'_{1_p}x'_2$
7. Value equivalence	$\underline{x_{1_p}}x_2 \sim x'_{1_p}x'_2$

Note. Classification introduced in Farquhar (1984, p. 1285). The underlined attribute is the one that changes during the elicitation process.

utility, the whole risk attitude is captured by the shape of the utility function u. If the utility function is concave, the individual is risk-averse; if it is convex, the individual is risk-seeking; and if it is linear, the individual is risk neutral. To measure risk attitudes, it is thus sufficient to know the shape of the utility function. In economics, this is often summarised by the Arrow–Pratt index $A(x) = -u''(x)/u'(x)$, which accounts for curvature independently of the scale of the utility function. The standard hypothesis is that the Arrow–Pratt index (i.e. risk aversion) falls with wealth (see Eeckhoudt et al., 2005, for a detailed presentation of expected utility theory).

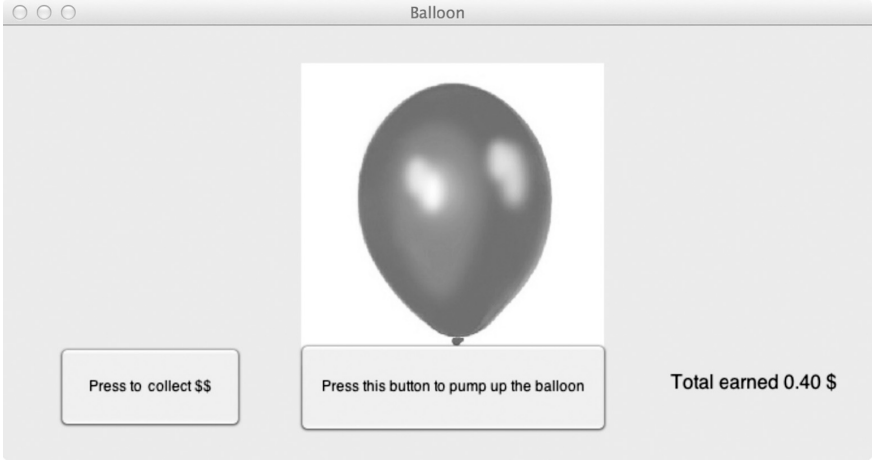
7.4.2 Elicitation Methods for Risk Attitudes

Farquhar (1984) distinguishes two broad categories of preference-elicitation methods. The first consists of comparisons between a sure amount and a lottery; this is the *standard gamble*. In the second, called *paired gamble*, elicitation is based on comparisons between two non-degenerate lotteries. The classification is further described in Table 7.9, where the underlined attribute is the elicitation variable, i.e. the variable that changes during the elicitation process. For both the standard gamble and paired gamble, elicitation is based on either a single comparison between lotteries or an equivalence between lotteries. In the former (comparison) the choice is binary, while in the latter (equivalence) the choice corresponds to the specification of an indifference value.

Starting with Binswanger (1980), one tradition in experimental economics is to elicit risk preferences using a series of choices between abstract lotteries. This corresponds to preference comparison in Table 7.9. A typical example is Hey and Orme (1994), who mix standard-gamble and paired-gamble preference-comparison methods to estimate various decision models. Another tradition follows the usual theoretical approach in decision under risk and measures certainty equivalents through indifference. Good examples of this measurement of risk preferences using certainty equivalence can be found in Becker and Brownson (1964) and Tversky and Kahneman (1992). Alternative methods have been proposed in the literature, based in particular on balloon tasks (Lejuez et al., 2002) (see Focus 7.16) or investment choices (see Focus 7.17).

Focus 7.16
The balloon analogue risk task (BART)

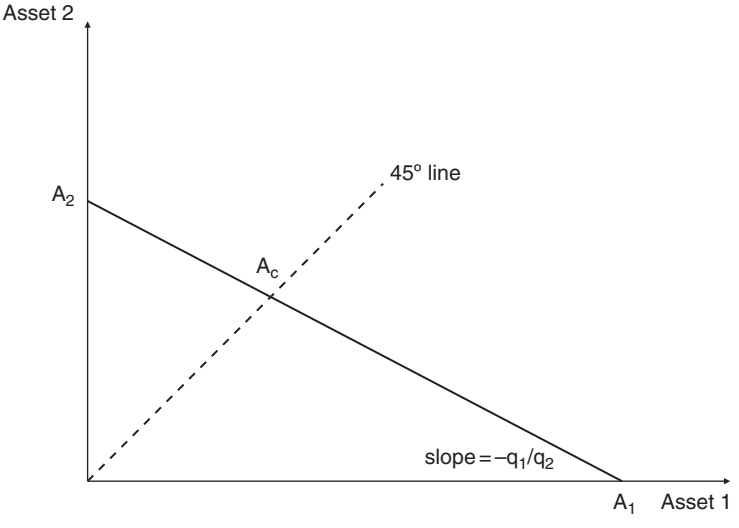
The balloon analogue risk task (Lejuez et al., 2002) is an intuitive task designed to measure risk attitudes without the usual numerical representation of lotteries. The task is simple: the participant is presented with a balloon into which the subject can choose to pump air. Subjects earn money each time they pump air into the balloon. At the same time, as long as the balloon does not explode, each pump causes the balloon to inflate. At any moment, a subject can choose to cash out prior earnings. If the balloon explodes, the earnings from pumping are lost. Participants are not informed about the balloon's breaking point. In BART, pumping corresponds to sampling without replacement from an unknown urn in which one ball out of n causes an explosion. Assuming that subjects know the number of balls n, it is possible to represent each pumping choice $j = 1, \ldots n$ as a binary lottery paying 0 with probability $\frac{1}{n-j-1}$ and paying cumulated earnings (j times the money increment) with probability $\frac{n-2-j}{n-j-1}$.



Crosetto and Filippin (2016a) compare the BART with a paired-gamble method (the Holt and Laury (2002), method) and a survey question on risk attitudes in general (Dohmen et al., 2011). The results show that while risk aversion shows up when preferences are elicited based on the Holt and Laury (2002) method (as well as in alternative elicitation methods), BART reveals considerable risk loving. The BART is, however, positively and moderately correlated with the answer to the survey question on risk attitudes. Crosetto and Filippin (2006) also point out that the measure of preference parameters with BART is difficult. First, a risk-aversion parameter can be calculated only under the assumption that the subjects know the details of the underlying urn, or are able to form precise beliefs about its composition. Second, as the explosion of the balloon is random, the stopping point is not observed when the balloon explodes.

Focus 7.17
Portfolio choice and the elicitation of risk attitudes

The investment decision in Gneezy and Potters (1997) and Charness and Kuhn (2010) is an alternative to the usual preference-elicitation methods. Instead of using abstract choices between lotteries, the method replicates standard portfolio-selection choice. The subject is endowed with a given amount of money that they can allocate between a risky asset and a risk-free asset. The risky asset pays $1 + rr$ times the amount invested with probability p and nothing otherwise. The money invested in the risk-free asset is kept by the subject. The rate of return r is such that the expected rate of return on the risky asset is strictly higher than the rate of return on the safe asset: $p(1 + rr) > 1$. The equity premium for the risky asset is strictly positive and any risk-neutral or risk-loving subject should invest all the endowment in the risky asset. In addition, the greater the subject's risk aversion, the less should be the investment in the risky asset. Charness and Kuhn (2010) use $rr = 1.5$ and $p = 0.5$. Charness and Viceisza (2012) find the method to have more predictive power than the Holt and Laury (2002) method, using rural subjects in Senegal. Choi et al. (2007) also use portfolio choices to elicit risk attitudes in the lab. Each subject is endowed with a given amount of money that they can allocate between two stage-contingent securities, to be bought at given prices. Each security offers a payoff of one unit in one state and nothing in the other. The price of asset 1 is q_1 and that of asset 2 is q_2. Subjects see the budget constraint associated with their portfolio allocation choice and select their preferred point on the budget constraint. As shown in the figure below, certainty corresponds to the intersection between the budget constraint and the 45° line: half of the endowment is invested in asset 1 and half in asset 2 (point A_c).



The polar cases are when subjects allocate all their endowment to either asset 1 (point A_1) or asset 2 (point A_2). Choi et al. (2007) use 50 different budget lines in their experiment. The experimental design allows them to test the general axiom of revealed preferences and to estimate risk-aversion and preference parameters by either maximum likelihood or non-linear least squares. In Choi et al. (2007) a large majority of choices satisfy the general axiom of revealed preference and a majority of subjects have kinked non-expected utility indifference curves that are consistent with loss or disappointment aversion.

Probability-Equivalence Versus Certainty-Equivalence Methods

A large literature in decision analysis compares the probability-equivalence (PE) and certainty-equivalence (CE) methods, and concludes that certainty equivalence should usually be preferred to probability equivalence. The typical finding is that PE produces higher utilities, i.e. more risk aversion, than CE (Hershey et al., 1982; Hershey and Schoemaker, 1985; Slovic et al., 1990; Delquié, 1993). Johnson (1984) suggests that the difference between the two methods is not entirely due to framing, but could also be related to the response mechanism. Hershey and Schoemaker (1985) conjecture that subjects use the certain outcome as a reference point when answering in the PE method. However, Bleichrodt (2001) shows that a quantitative assessment of this conjecture is fairly complex as the subject's reference point varies between the PE questions. In this case, subjects behave as if they were facing lotteries with gains and losses, and loss aversion can explain the increase in risk aversion. Bleichrodt (2001) proposes a solution to this difficulty.

On the other hand, CE yields responses that are biased towards the expected value, and sometimes towards risk seeking (Morrison, 2000). The difference in utilities in the PE and CE methods poses a serious descriptive challenge to expected-utility theory. Under expected utility, both methods should produce the same utility, i.e. the same measure of individual risk aversion. Elicitation procedures that are equivalent in theory should yield identical results. In addition, Karmarkar (1978) and McCord and Neufville (1985) showed that the utility function elicited via CE depends on the probabilities used: a more concave utility function is produced with higher probabilities. Machina (1987) called this the *utility-evaluation effect*, which can be explained by rank-dependent utility theory (Quiggin, 1982) or prospect theory (Kahneman and Tversky, 1979). These latter are based on the non-linear transformation of probabilities, and risk aversion that varies with the probability level (in addition to loss aversion, the PE method is also distorted by probability transformation (Bleichrodt et al., 2001)). The utility-evaluation effect is also compatible with hypothesis II in Machina (1982). The differences observed between PE and CE disqualify expected utility as a descriptive theory of choice under risk. Standard PE or CE leads to systematic bias in utility measurement, and therefore in the assessment of risk attitudes. In light of this result, we may ask whether the use of PE or CE makes sense. Bleichrodt et al. (2001) show that CE and PE can still be used if they are appropriately corrected for the main sources of bias, probability transformation and loss aversion. Bleichrodt et al. (2001) propose various methods for this correction. In particular, the use of a probability level close to $1/3$ in the CE method helps considerably in reducing the gap between the different elicitation methods.

Eliciting Indifference in Choice Under Risk

There are three main elicitation procedures to elicit indifferences. The first, *matching*, asks subjects to directly assess their indifference point. The second, *bisection*, determines the indifference point through a convergent series of choices. The third, *bracketing*, infers indifference through a series of choices.

An example of a matching task, based on a standard gamble method for certainty equivalence, is the following:

Table 7.10 An example of the bisection procedure

Iteration	Choices offered in the elicitation procedure		
1	**10**	vs.	$20_{0.5}0$
2	5	vs.	$\mathbf{20_{0.5}0}$
3	**7.50**	vs.	$20_{0.5}0$
4	6.25	vs.	$\mathbf{20_{0.5}0}$
5	**6.87**	vs.	$20_{0.5}0$
Indifference value	6.56		

Table 7.11 The bracketing procedure in Tversky and Kahneman (1992)

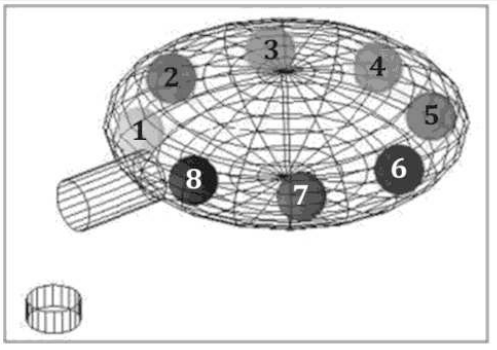
Choice	Alternative A	Alternative B
1	$150_{0.25}50$	**150**
2	$150_{0.25}50$	**125**
3	$150_{0.25}50$	**104**
4	$150_{0.25}50$	**87**
5	$150_{0.25}50$	**72**
6	$\mathbf{150_{0.25}50}$	60
7	$\mathbf{150_{0.25}50}$	50
Indifference value		66

What amount of money, \$ __ , if paid to you with certainty, would make you indifferent to the lottery paying \$$x_1$ with probability p and x_2 with probability $1 - p$.

In the matching task, participants are asked to state their certainty-equivalent value for the lottery $x_1{}_p x_2$. In the literature, the matching procedure is considered to be dominated by choice-based approaches, such as bisection or bracketing, as inferring indifference from a series of choices leads to fewer inconsistencies and less noise than asking subjects directly for their indifference values (Bostic et al., 1990; Hey et al., 2009).

Abdellaoui et al. (2008) elicit certainty equivalents using a second elicitation procedure: bisection. In each choice a subject is faced with two lotteries: one fixed risky lottery and another that is always riskless. Table 7.10 shows an example of the bisection procedure, with a risky lottery offering $20_{.5}0$. For each iteration, the lottery that is chosen is printed in bold. In the example, the starting values in the iterations are chosen so that the two lotteries are of equal expected value. Different starting values can be chosen. Depending on the choice made, the certain outcome is increased or decreased. In the example, the size of the change is always half the size of the change in the previous question. In Abdellaoui et al. (2008), the numbers are always integers. In Table 7.10, the numbers have two digits after the decimal point. The bisection procedure yields an interval in which the indifference value should lie, the midpoint of which is taken as the indifference value.

Tversky and Kahneman (1992) use the third procedure of bracketing to determine certainty equivalents. Subjects make a series of seven choices between a fixed risky lottery and riskless lotteries. Table 7.11 shows a series of bracketing choices, with the lottery as alternative A and the certain amount as alternative B. Subjects choose between



Win **€10** if one of the following balls is extracted:

1 2 3 4

Win **€0** if one of the following balls is extracted:

5 6 7 8

Lottery	Sure amount	
O	O	€0.50 for sure
O	O	€1.00 for sure
O	O	€1.50 for sure
O	O	€2.00 for sure
O	O	€2.50 for sure
O	O	€3.00 for sure
O	O	€3.50 for sure
O	O	€4.00 for sure
O	O	€4.50 for sure
O	O	€5.00 for sure
O	O	€5.50 for sure
O	O	€6.00 for sure
O	O	€6.50 for sure
O	O	€7.00 for sure
O	O	€7.50 for sure
O	O	€8.00 for sure
O	O	€8.50 for sure
O	O	€9.00 for sure
O	O	€9.50 for sure

Figure 7.13 The bracketing procedure used in L'Haridon and Vieider (2015)

A and B. The certain amounts are linearly spaced on the log scale, as in Tversky and Kahneman (1992). In Table 7.11 the lottery that is chosen is printed in bold. Bracketing yields an interval within which the indifference value should lie. The mid-point of this interval is taken as the indifference value.[19] In Table 7.11 this lies between 72 and 60, and is taken to be 66. Different scales can be used to build the choices. As an example, Figure 7.13 shows an example of bracketing with linearly spaced riskless alternatives, similar to that in Cohen et al. (1987).

There are two differences between bracketing and bisection. First, the bracketing procedure is not chained: each choice in Table 7.11 is independent of the subject's previous answers. Second, for a given number of choices, bracketing is less precise than bisection. As a result, the choice between the methods trades off precision and chaining: if the propagation of errors is an issue in the experiment, bracketing should be preferred; if precision is a concern, bisection should be chosen.

7.4.3 A Simple Visual Method: Binswanger (1980)

One of the first methods for eliciting choice under risk was proposed by Binswanger (1980). Subjects have only one choice to make. The method is quick, easy to implement

[19] In the case of multiple switching, this is, of course, no longer the case. Multiple switching issues will be addressed in our presentation of the Holt and Laury (2002) method.

Table 7.12 The payoffs and risk classification in Binswanger (1980)

Choice	Low payoff (heads)	High payoff (tails)	Risk-aversion class	CRRA index interval
O	50	50	Extreme	$]7.51; \infty[$
A	45	95	Severe	$]1.74; 7.51[$
B	40	120	Intermediate	$]0.812; 1.74[$
D*	35	125	Inefficient	
C	30	150	Moderate	$]0.316; 0.812[$
D	20	160	Inefficient	
E	10	190	Slight to neutral	$]0; 0.316[$
F	0	200	Neutral to negative	$]-\infty; 0[$

Table 7.13 The payoffs and risk classification in Eckel and Grossman (2008)

Choice	Low payoff (Heads)	High payoff (Tails)	Risk aversion class
1	16	16	Extreme
2	12	24	Intermediate
3	8	32	Moderate
4	4	40	Slight-to-neutral
5	0	48	Risk neutral

and easily understood. The original method appeared as part of an experiment in Indian villages. Subjects select a lottery from a set of possible lotteries. In the original work of Binswanger (1980), all possible binary 50–50 lotteries were included. Subjects are given a form describing eight lotteries, denoted O to F, and asked to choose one. Table 7.12 shows the payoff characteristics of the eight lotteries and the corresponding risk-aversion class and CRRA index interval. The variance and expected payoff rise from lottery O to lottery F – lotteries B and D*, C and D, E and F have the same expected payoffs.

Subjects face a sequence of lotteries with different payoffs (multiples of the payoffs in Table 7.12 by factors of 1/100, 1/10, 1 and 10). When the payoffs are small, about half the respondents correspond to intermediate and moderate risk-aversion preferences. Nearly a third are close to risk-neutral or risk-loving, and under 10% were highly risk-averse. However, as payoffs rise, nearly 80% of subjects display moderate risk aversion, with risk-neutral and risk-loving behaviour almost disappearing. The standard prediction on the link between risk aversion and wealth is thus supported: subjects' willingness to accept small bets of a fixed size increases with wealth. This corresponds to decreasing absolute risk aversion (DARA).[20]

Eckel and Grossman (2008) offer a popular implementation of the Binswanger (1980) method based on a linearly increasing expected value and greater standard deviation in the sequence of lotteries. Table 7.13 shows the payoff characteristics of the five lotteries used in this implementation.

Cardenas and Carpenter (2013) use a variant of the lotteries O, A, B, C, E and F in Binswanger (1980) to measure risk attitudes in a representative sample of 3,000 subjects

[20] Subjects also display increasing, and then decreasing, relative risk aversion: after an initial increase, aversion to multiplicative risks (i.e. risks expressed as a percentage of wealth) fall with wealth.

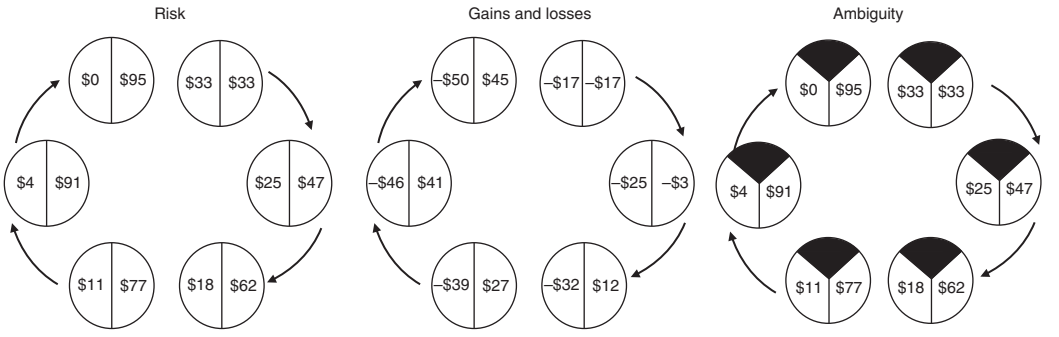


Figure 7.14 The Binswanger (1980) method in Carpenter and Cadernas (2013)

in six Latin American cities. They also adapt the method to include risk pooling, attitudes towards gains and losses, and ambiguity, as well as a measure of well-being. The elicitation task is displayed in graphical form, as shown Figure 7.14. In the ambiguity treatment, the odds of a high or low payment were bounded between 3/10 and 7/10, but are unknown. This uncertainty about probability is represented by the black areas in Figure 7.14.

7.4.4 A Well-Known Paired-Gamble Method: Holt and Laury (2002)

The Holt and Laury (2002) method is based on an intuitive, simple design, which is widely used to elicit risk aversion (see Cohen et al., 1987, for an earlier example based on certainty equivalence). Each subject is presented with a menu of ten ordered choices between paired lotteries. Each pair of lotteries consists of a single choice between a safe option (option A) and a risky option (option B) of the following form:

$$x_{A1_p}x_{A2} \text{ vs. } x_{B1_p}x_{B2}$$

with x_{A1} greater than x_{A2} and x_{B1} greater than x_{B2}. Option A is safe in the sense that it offers less variable payoffs than option B: $x_{A1} < x_{B1}$ and $x_{A2} > x_{B2}$.

Description of the Method

More precisely, subjects make ten different choices ordered by increasing probability p from 0.1 to 1. The expected payoff difference between options A and B,

$$p[(x_{A1} - x_{A2}) - (x_{B1} - x_{B2})] + (x_{A2} - x_{B2})$$

falls with p and is negative for

$$p > (x_{A2} - x_{B2})/[(x_{B1} - x_{B2}) - (x_{A1} - x_{A2})]$$

Regarding risk attitudes, only extreme risk seekers would choose option B at any probability level p, even when this is close to 0. Symmetrically, only an extreme risk-averse decision-maker would keep on choosing option A for any probability p less than 1. Note that no one should choose the dominated option A when the probability p is 1. A risk-neutral individual would switch from option A to option B when $p > (x_{A2} - x_{B2})/[(x_{B1} - x_{B2}) - (x_{A1} - x_{A2})]$. Here the point at which the decision-maker

Table 7.14 The ten paired lottery-choice decisions in Holt and Laury (2002)

Choice number	Option A	Option B	Expected payoff ≠
1	1/10 of 2, 9/10 of 1.6	1/10 of 3.85, 9/10 of 0.1	+1.17
2	2/10 of 2, 8/10 of 1.6	2/10 of 3.85, 8/10 of 0.1	+0.83
3	3/10 of 2, 7/10 of 1.6	3/10 of 3.85, 7/10 of 0.1	+0.50
4	4/10 of 2, 6/10 of 1.6	4/10 of 3.85, 6/10 of 0.1	+0.16
5	5/10 of 2, 5/10 of 1.6	5/10 of 3.85, 5/10 of 0.1	−0.18
6	6/10 of 2, 4/10 of 1.6	6/10 of 3.85, 4/10 of 0.1	−0.51
7	7/10 of 2, 3/10 of 1.6	7/10 of 3.85, 3/10 of 0.1	−0.85
8	8/10 of 2, 2/10 of 1.6	8/10 of 3.85, 2/10 of 0.1	−1.18
9	9/10 of 2, 1/10 of 1.6	9/10 of 3.85, 1/10 of 0.1	−1.52
10	10/10 of 2, 0/10 of 1.6	10/10 of 3.85, 0/10 of 0.1	−1.85

Table 7.15 Lottery-choice decisions and the CRRA index

Switching choice	CRRA index interval
1	$]-\infty; -0.95[$
2	$]-0.95; -0.49[$
3	$]-0.49; -0.15[$
4	$]-0.15; 0.15[$
5	$]0.15; 0.41[$
6	$]0.41; 0.68[$
7	$]0.68; 0.97[$
8	$]0.97; 1.37[$
9	$]1.37; \infty[$

shifts from option A to option B reveals their risk attitude. Table 7.14 shows the choices that the subjects face when $x_{A1} = 2$, $x_{A2} = 1.6$, $x_{B1} = 3.85$ and $x_{B2} = 5$.

Subjects who switch from A to B between choices 4 and 5 are risk-neutral, while those who switch between choices 2 and 3 are significantly risk-seeking and those switching between choices 7 and 10 are significantly risk-averse.

A major strength of eliciting indifferences is that it allows the direct estimation of individual relative risk aversion based on a particular utility function. Consider, for instance, a particular specification of a constant relative risk-aversion (CRRA) utility function,

$$u(x) = \begin{cases} x^{1-\theta}/(1-\theta) & \text{if } \theta \neq 1 \\ log(x) & \text{if } \theta = 1 \end{cases} \tag{7.2}$$

The shape of the individual utility can be inferred from subject's choices between options A and B.

An expected-utility maximiser with risk-aversion parameter θ is indifferent between the two options if $p \times 2^{1-\theta} + (1-p) \times 1.6^{1-\theta} = p \times 3.85^{1-\theta} + (1-p) \times 5^{1-\theta}$. Solving this equation for θ yields the degree of risk aversion as a function of the probability p at which a subject switches between options A and B. This switching point produces a direct measure of the CRRA index interval, as in Table 7.15. A subject switching between choices 4 and 5 is close to risk-neutral. It is not clear, however, where the subject's preference lies in this interval: the subject may be slightly risk-averse, slightly risk-seeking or risk neutral.

Focus 7.18
Incentives and repeated choice

The Holt and Laury (2002) experiment contains an original feature in designing incentives for repeated choice. Subjects face three repeated treatments: one hypothetical treatment and two treatments with incentives. Among the treatments involving real incentives, subjects first face a low-payoff treatment, as in Table 7.14, and then a high-payoff treatment with the payoffs multiplied by 20. Subjects are paid for real for both tasks, with the outcome of the low-payoff task determined before the high-payoff task begins. This procedure is usually problematic as it creates wealth effects between the repeated treatments. The common way of controlling for wealth effects in repeated choice is to pay only one task, selected at random at the end of the experiment. This procedure is often criticised as incentives are diluted by randomisation (see Chapter 5, Section 5.2.3, for a discussion). The question of what is the real choice under randomisation is also unclear. Holt and Laury (2002) take an alternative approach and introduce a procedure for controlling for wealth effects while paying for real for both tasks. To control for wealth effects, subjects are asked to give up their earnings in the low-payoff task before carrying out the high-payoff task. If subjects wish to participate in the high-payoff treatment they have to pay back their experimental earnings from the low-payoff task. In the experiment, no subject declined to participate and the design succeeded in paying each subject for both choices. An open question remains regarding the behavioural consequences of paying back previously earned outcomes. While controlling for wealth effects, the procedure does not control for the behavioural consequences of paying something to the experimenter in the middle of the experiment. Framing and status quo effects may be at play and influence behaviour in the high-payoff treatment.

Experimental Evidence

The Holt and Laury (2002) method is widely cited in the literature and commonly used in experimental economics (see Holt and Sherman, 2014, for a survey of applications). Its popularity can first be attributed to its capacity to deal in a simple way with incentive compatibility in experiments, including choice under risk, and to its findings corresponding to the standard presumptions on the effect of incentives.[21] Second, the method is highly tractable: only one table is used to obtain an indicator of risk aversion based on the number of A choices. It can also be used either with a computer or a paper-and-pencil questionnaire. It is similar in this respect to that in Tanaka et al. (2010), and provides a simple assessment of risk attitudes at low cost. Many lab experiments indeed use the method to control for risk attitudes by adding a table like Table 7.15 to the end of experiments on public-good games, auctions, experimental markets, etc.

The method has been implemented in a number of ways (Harrison and Rutström, 2008), including different price lists with or without iteration (Andersen et al., 2009) and binary choices *à la* Hey and Orme (1994). In the former, subjects are presented with the entire list of prices and choose sequentially between them, while in the latter subjects are faced with ten independent lotteries. Aside from order effects (Harrison

[21] The original paper also contains an original design for incentives under repeated choice, described in Focus 7.18.

et al., 2005), a number of papers have focused on the effect that the particular frame in Holt and Laury's price list might have on risk preferences. For example, Andersen et al. (2009) modify the linear nature of the Holt and Laury list and introduce skewed lists. The first treatment, called the SKEWHI treatment, offers a list of probabilities, $p = (0.3, 0.5, 0.7, 0.8, 0.9, 1)$. The second treatment, called SKEWLO, offers probabilities $p = (0.1, 0.2, 0.3, 0.5, 0.7, 1)$. While SKEWHI leads to the same power parameter (under expected utility), SKEWLO increases it. Bosch-Domènech and Silvestre (2013) change the frame by removing some pairs of gambles. The removal of the pairs with the highest expected value (at the end of the list) reduces risk aversion. If the list contains a certain amount (and then elicits a certainty equivalent), no effect of removal was found. The authors also find no role for the way the list is ordered. Lévy-Garboua et al. (2012) found opposite results, with a large impact of the order of appearance of the ten lottery pairs.

All previous discussions are conditional on the implicit assumption that decision-makers are expected-utility maximisers. If the decision-maker has non-expected-utility preferences, the transformation of probabilities can bias the assessment of utilities. In non-expected-utility decision models, risk aversion depends on the probability, which increases from the top to the bottom of the decision table. In other words, the utility function, which is measured over the outcome scale, will be dependent on the (probability) scale used to measure it. A subject who overweights low probabilities and underweights high probabilities may choose option B at low probability and option A at high probability. Drichoutis and Lusk (2016) show that probability weighting is a source of bias in the Holt and Laury method. They propose a new method in which the probability remains constant across the ten decision tasks in the Holt and Laury table, while keeping the CRRA utility parameter implied by a switch at a given row (see Table 7.15).

As a control, Drichoutis and Lusk (2016) also consider a table where the matched utility parameter interval given by Table 7.15 is compatible with a one-parameter probability weighting function with a parameter value equal to 0.6.[22] To be consistent with the original Holt and Laury experiment, in which risk aversion is measured for different stakes, they also propose a treatment where all amounts are scaled up by a factor of five. In the design of Drichoutis and Lusk (2016), all rows are presented to subjects separately as binary choices. Different treatments (especially treatments with different scales of payoff) are presented in random order. The authors find that the HL method with constant probability described in Table 7.16 increases the elicited power coefficient for low payoffs (power = 1.11) but not for high payoffs (power = 0.2). Their estimate of the probability-weighting parameter (based on probability 0.5 and a one-parameter probability weighting function) is 3.1, which is clearly out of the common range of estimated values found in the literature (see Booij et al., 2010, Table 1, for a survey). Focus 7.19 reviews the literature on the comparison between the Holt and Laury method and certainty-equivalence method. Focus 7.20 shows how incentivised elicitation compares to self-reported risk preferences.

[22] Probability weighting is defined in Section 7.4.5, Focus 7.25.

Table 7.16 The ten paired lottery-choice decisions in Drichoutis and Lusk (2016)

Choice number	Option A	Option B	Expected payoff ≠
1	5/10 of 1.68, 5/10 of 1.6	5/10 of 2.01, 5/10 of 1.00	+0.13
2	5/10 of 1.76, 5/10 of 1.6	5/10 of 2.17, 5/10 of 1.00	+0.10
3	5/10 of 1.84, 5/10 of 1.6	5/10 of 2.32, 5/10 of 1.00	+0.06
4	5/10 of 1.92, 5/10 of 1.6	5/10 of 2.48, 5/10 of 1.00	+0.02
5	5/10 of 2.00, 5/10 of 1.6	5/10 of 2.65, 5/10 of 1.00	−0.03
6	5/10 of 2.08, 5/10 of 1.6	5/10 of 2.86, 5/10 of 1.00	−0.09
7	5/10 of 2.16, 5/10 of 1.6	5/10 of 3.14, 5/10 of 1.00	−0.19
8	5/10 of 2.24, 5/10 of 1.6	5/10 of 3.54, 5/10 of 1.00	−0.35
9	5/10 of 2.32, 5/10 of 1.6	5/10 of 4.50, 5/10 of 1.00	−0.79
10	5/10 of 2.40, 5/10 of 1.6	5/10 of 4.70, 5/10 of 1.00	−0.85

Estimating Risk Preferences

The 10 paired-lottery choices in Holt and Laury (2002) not only allow for the classification of individuals based on their switching point between the two options and/or the calculation of the bounds implied by this switching point. As shown by Holt and Laury (2002) and popularised by Harrison and Rutström (2008), the method also allows the estimation of the structural decision model of choice by maximum likelihood, in the spirit of Sopher and Gigliotti (1993) and Hey and Orme (1994).

Assume that the choice data between options A and B for a given choice j are determined by a given data-generating process. Denote Z_j the dummy variable for option A being selected in choice number j in Table 7.14. $U(A_j)$ and $U(B_j)$ denote the value associated with the prospect offered by option A and B in task j. We focus on expected utility and assume that $U(.)$ is defined by (7.1).[23] The deterministic decision rule for an expected-utility maximiser is therefore

$$DR_j = U(A_j) - U(B_j) \tag{7.3}$$

Under this rule, the decision-maker chooses A over B in choice j if $DR_j > 0$, chooses B over A if $DR_j < 0$, and is indifferent between them if $DR_j = 0$. If the utility function underlying choices is further assumed to match the specification in (7.2), the decision rule is parametrised by the preference parameter θ, which is the CRRA index reflecting (relative) risk aversion:

$$DR_j(\theta) = U(A_j|\theta) - U(B_j|\theta) \tag{7.4}$$

This decision rule predicts all the decision-maker's choices between options A and B when the preference parameter is θ. The model can be extended to the assumption that

[23] The analysis can be generalised to non-expected-utility theory; see Moffatt (2015) for a complete overview of preference function estimation under risk.

Focus 7.19
Comparing standard-gamble methods

Anderson and Mellor (2009) compare the measures obtained from two standard-gamble elicitation methods: the Holt and Laury and certainty-equivalence methods. The certainty-equivalence method is based on two series of hypothetical tasks between risky and safe options – one between a job with either a certain or a risky level of income, and the other between inheriting either a certain or a risky level of wealth. In the Holt and Laury task, the safe options involve payoffs of $6.00 and $4.80, and the risky options payoffs of $11.55 and $0.30. The survey measure is based on the following question, taken from a well-known survey (the Health and Retirement Study – Barsky et al., 1997):

Suppose that you are the only income earner in the family. Your doctor recommends that you move because of allergies, and you have to choose between two possible jobs. The first would guarantee you an annual income for life that is equal to your parents' current total family income. The second is possibly better-paying, but the income is also less certain. There is a 50–50 chance the second job would double your total lifetime income and a 50–50 chance that it would cut it by a third. Which job would you take: the first job or the second job?

There are three possible answers: First job/Second job/Do not know. The survey also includes four additional questions where the downside risk associated with the second job is equal to a 50%, a 75%, a 20% and a 10% cut in outcome. The second choice has the same characteristics in terms of probability and downside risk. The measure is based on the following question:

Suppose that a distant relative left you a share in a private business worth one million dollars. You are immediately faced with a choice whether to cash out now and take the one million dollars, or to wait until the company goes public in one month, which would give you a 50–50 chance of doubling your money to two million dollars and a 50–50 chance of losing one-third of it, leaving you 667 thousand dollars. Would you cash out immediately or wait until after the company goes public?

There are three possible answers: Cash out/Wait/Do not know. Based on the answers to the five questions, subjects are classified into one of six categories ranging from least risk-tolerant (rejecting the new risky option when the downside risk is a loss of 33%, 20% and 10%) to most risk-tolerant (accepting the risky option when the downside risk is a loss of 33%, 50% and 75%). Some 239 subjects carried out both the Holt and Laury probability-equivalence task and the certainty-equivalence tasks. Anderson and Mellor (2009) find no systematic correlation between the risk aversion from the certainty-equivalence task based on job gambles and that in the probability-equivalence task. For example, restricting the analysis to the 97 subjects who answered the survey in a consistent manner, the correlation between the experimental measure and the survey-based classification is only 0.16; the correlation association between the experimental measure and the classification from the hypothetical inheritance questions was only slightly higher at 0.22.

individuals implement their intentions with some error (Sopher and Gigliotti, 1993; Hey, 2005). This error, denoted ϵ, reflects anything that prevents the decision-maker from behaving according to the deterministic-choice rule and its underlying decision model: misperceptions, miscalculations or inattention. In this case, the decision-maker chooses option A_j over option B_j if $DR_j(\theta)+\epsilon_j > 0$. Or, alternatively, the decision-maker chooses

Focus 7.20
Survey questions and the measurement of risk attitudes

Dohmen et al. (2011) compare self-reported risk attitudes to incentivised experimental measures. The first source of data is a large representative survey of the adult German population, in which respondents are asked:

How willing are you to take risks, in general? Rate your willingness on a scale between 0 and 10

The question is also asked in less general contexts, referring to risk attitudes in car driving, financial matters, sport, work, health and trust in others. The second source of data is an experiment run on a pool of 450 subjects that is representative of the adult German population. The experiment includes two parts: subjects are first asked to answer a questionnaire, which includes the simple question quoted above. Subjects then play a real-stakes lottery experiment, framed in the same way as in the experiment shown in Figure 7.13: subjects have to make a series of 20 choices presented in ascending order between a sure option and a lottery $300_{0.5}0$. The sure options range from 0 to 190 euros. The switching point between the lottery and the sure option defines the certainty equivalent for the lottery $300_{0.5}0$. The distance between the certainty equivalent and the expected value of the lottery (150 euros) measures the intensity of (weak) risk aversion for each individual by their risk premium. The incentive compatibility of the experiment was ensured by a between-subject random-task incentive system: each subject had a $^1/_7$ choice to have one of their choices played for real. The median (mean) response to the general risk question in the large panel is 5 (4.42), with a standard deviation of 2.38, reflecting considerable heterogeneity in risk attitudes. The answers to the general question in the separate experimental pool were similar. Moreover, 78% of these subjects provided a certainty equivalent indicating risk aversion. To check whether survey data can predict risk attitudes in the real-stakes experiment, Dohmen et al. (2011) first regressed the value of the safe option at the switching point on the answers given to the general risk question, producing a significant estimated coefficient of 0.61. Once biological, socio-economic and survey-condition controls were included, this coefficient remained significant but fell to 0.40. In a similar experiment, using the Holt and Laury (2002) method instead of the certainty-equivalence method, Attanasi et al. (2017) find a similar correlation of 0.47 between the self-reported risk attitude and the risk ordering in the decision task. Eliciting risk attitudes with four incentivised risk elicitation tasks (the Holt and Laury (2002) method, the Eckel and Grossman (2008), the investment game in Gneezy and Potters (1997) and their own elicitation task) and two survey measures, Crosetto and Filippin (2016b) found correlations ranging from 0.03 to 0.30 between incentivized tasks and the survey measures. Dohmen et al. (2011) also observe consistency in the answers to the general and domain-specific risk questions. A principal-components analysis reveals that 60% of the variation in individual risk attitudes can be explained by one principal component, with factor loadings for the different risk question ranging from 0.74 to 0.81. These findings are consistent with the existence of a single underlying behavioural trait determining risk attitudes. In a large sample experiment with nearly 3,000 subjects over 30 countries, Vieider et al. (2015) show this finding holds in most countries, with incentivised measures for gains and losses and for risk and uncertainty. In addition, Vieider et al. (2015) find incentivised and survey measures to correlate between countries.

B over A in choice j, and contradicts their deterministic decision rule, if the error is large enough: $\epsilon_j > -DR_j(\theta)$. Holt and Laury (2002) make use of the stochastic process in Luce (1959), so that the probability that the decision-maker selects A for choice j is

$$Pr\left[A_j|\theta,\eta\right] = \frac{U(A_j|\theta)^{1/\eta}}{U(A_j|\theta)^{1/\eta} + U(B_j|\theta)^{1/\eta}} \tag{7.5}$$

where η denotes the noise, or error, parameter. As the noise parameter η tends to 0, the choice becomes deterministic and $Pr\left[A_j|\theta,\eta\right]$ tends to 1. On the contrary, as the noise parameter becomes large, subjects make their choice at random between the two options and $Pr\left[A_j|\theta,\eta\right]$ tends to one-half. Equation (7.5) is the likelihood contribution for a single subject's decision in choice j for parameters θ and η.[24] For a full set of decisions $j = 1,\ldots,J$, the likelihood associated with a given subject's choices is

$$L(\theta,\eta) = \prod_{j=1}^{J} Pr\left[A_j|\theta,\eta\right]^{Z_j} \times \left(1 - Pr\left[A_j|\theta,\eta\right]\right)^{1-Z_j} \tag{7.6}$$

where $Z_j = 1$ if the subject chooses option A in choice j (and 0 otherwise).

Holt and Laury (2002) estimate an aggregate risk-preference parameter and an aggregate noise parameter. To obtain these aggregate estimates, the likelihood needs to be calculated over the N subjects in the experiment. The log likelihood to be maximised with respect to parameters θ and η is then

$$LL(\theta,\eta) = \sum_{i=1}^{N}\sum_{j=1}^{J} Z_{ij}\log\left(Pr\left[A_j|\theta,\eta\right]\right) + (1 - Z_{ij})\log\left(1 - Pr\left[A_j|\theta,\eta\right]\right)$$

where $Z_{ij} = 1$ if individual i chooses A in choice j (and 0 otherwise).

7.4.5 A Value-Equivalence Method: The Trade-Off Method

Wakker and Deneffe (1996) introduce a value-equivalence method – the 'trade-off method' – for eliciting utilities in decision under risk and uncertainty. This method is very general and is not restricted to the assumptions usually made in eliciting utility. Compared to most existing methods, this method is 'parameter-free' as it does not impose any parametric assumptions on the utility function. Moreover, it is not restricted to choice under risk and is easily applied to choice under uncertainty, in which the probabilities are unknown. Last, the trade-off method is not restricted to expected utility and can be used with a variety of preferences from expected utility to rank-dependent utility and prospect theory.

Description of the Method

In practice, the method elicits the utility function based on a 'standard sequence of outcomes' – a sequence of outcomes such that the utility difference between successive elements of the sequence is constant. It is based on inferences from indifferences

[24] If subjects are indifferent between the two options, the likelihood can be defined as the average of the likelihood of choosing either option; see Harrison and Rutström (2008).

between two two-outcome prospects. Consider, for example, $x_0 = \$10$ as the starting point of the standard sequence. First, an amount x_1 is elicited such that[25]

$$x_{10.5}1 \sim 10_{0.5}8 \tag{7.7}$$

The outcome x_1 can be elicited by a matching task or by a choice-based procedure. Most experiments use a choice-based procedure (Wakker and Deneffe, 1996; Abdellaoui, 2000). In the second experimental task, the amount x_1 is substituted for the outcome \$10 and an amount x_2 is elicited such that[26]

$$x_{20.5}1 \sim x_{10.5}8 \tag{7.8}$$

Together, (7.7) and (7.8) define the elements of a standard sequence. Under expected utility, the first indifference, in (7.7), leads to the following equality:

$$\frac{1}{2}u(10) + \frac{1}{2}u(8) = \frac{1}{2}u(x_1) + \frac{1}{2}u(1) \tag{7.9}$$

and the second, in (7.8), leads to

$$\frac{1}{2}u(x_1) + \frac{1}{2}u(8) = \frac{1}{2}u(x_2) + \frac{1}{2}u(1) \tag{7.10}$$

Subtracting one equation from the other and factoring out $\frac{1}{2}$, we find

$$u(x_2) - u(x_1) = u(x_1) - u(10) \tag{7.11}$$

which shows that the indifferences reveal an equality of utility differences between $u(x_2)$ and $u(x_1)$ and between $u(x_1)$ and $u(x_0)$ (with $x_0 = 10$). The amount x_1 is a mid-point (in utility terms) between x_0 and x_2. Figure 7.15.a shows how this utility mid-point is related to the curvature of the utility function. With a concave utility function, the utility mid-point x_1 is lower than the monetary mid-point $(x_0 + x_2)/2$. Moreover, the greater the difference, the more curved is the utility function, as shown by the grey area in Figure 7.15. On the contrary, with a convex utility function the utility mid-point is greater than the monetary mid-point. If the utility mid-point equals the monetary mid-point, then utility is linear – which corresponds to risk neutrality under expected utility.

A similar series of indifferences $x_{j+1_{0.5}}1 \sim x_{j_{0.5}}8$ can be elicited to obtain a longer sequence $x_0, x_1, x_2, \ldots, x_k$, where all outcomes are equally spaced in terms of utility

[25] In practice, the first step of the trade-off method is to select a starting outcome x_0 (\$10 in the example above) and to select outcomes A and a such that $A > a$ (8 and 1 in the example respectively).

[26] The answers are thus chained in the sense that the previous responses are used in the elicitation of subsequent choices. One issue with chaining is that it can lead to error propagation, where errors made in one particular choice affect later choices. Bleichrodt and Pinto (2000), Abdellaoui et al. (2005) and Bleichrodt et al. (2010) simulate error propagation and suggest that it is not a major concern in trade-off elicitations. Chaining can also lead to strategic answers. Bleichrodt et al. (2010) include a test of chaining by repeating the same two questions at the beginning of the experiment (when subjects did not yet know about the chaining of the answers) and towards the end (when they did). Were chaining to affect the results, the second set of questions should produce higher figures. Bleichrodt et al. (2010) find no evidence for this, and conclude that the chained nature of the measurements has no discernible impact on their results.

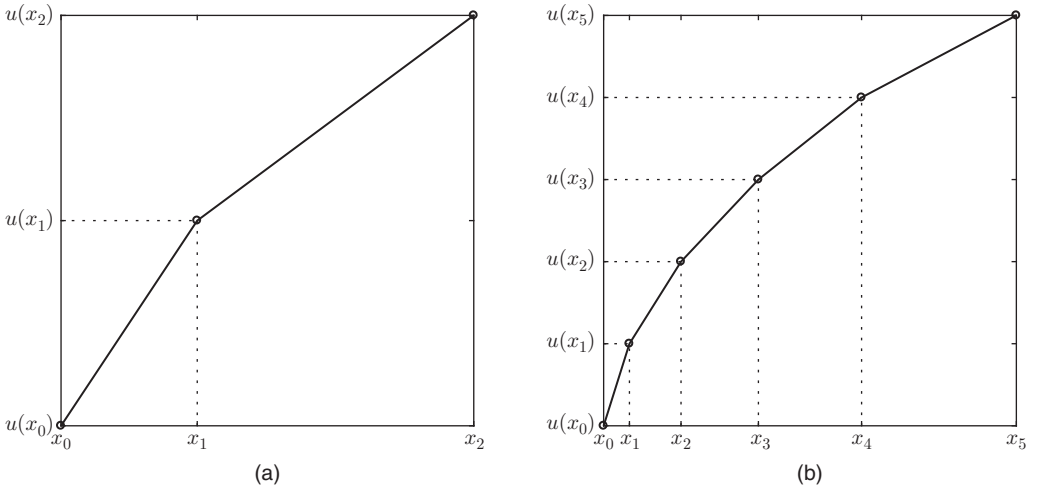


Figure 7.15 Trade-off sequences and elicited utility under risk and uncertainty

units. In this case, each utility value can be normalised to $u(x_j) = j/k$ by setting $u(x_0)$ equal to 0 and $u(x_k)$ equal to 1. Each element of the standard sequence corresponds to a direct observation of the inverse utility function, without any parametric assumption about the form of the utility function. This is illustrated via the six-element standard sequence in Figure 7.15.b. Focus 7.21 compares the trade-off methods with elicitations based on alternative methods.

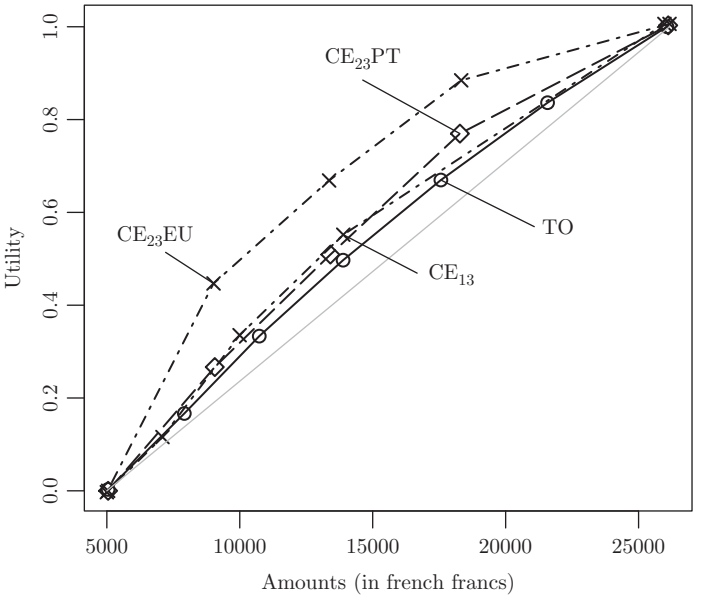
Measuring Utility Curvature

As shown in Figure 7.15, the area below the utility function provides a measure of the degree of concavity of the utility function. Most contributions apply the following method to calculate this area. Assume that the experiment has elicited a standard sequence $x_0, x_1, \ldots, x_5$. First, utility is normalised in order to take values between 0 and 1 by setting $u(x_0)$ equal to 0 and $u(x_5)$ to 1. Each element of the standard sequence defines a utility from 0 to 1 with increments of $1/5$. Second, the domain of utility is normalised to [0, 1], by transforming every outcome x_j to the value $(x_j - x_0)/(x_5 - x_0)$. Third, the normalised area under the utility function is calculated for each segment. Each subarea of a segment $x_{j+1} - x_j$ is the sum of a rectangle of area $(x_{j+1} - x_j) \times \frac{j}{5}, j = 1, \ldots, 4$ and a triangle of area $\frac{x_{j+1} - x_j}{2} \times \frac{1}{5}, j = 0, \ldots, 4$. If utility is linear, this area equals $1/2$. Utility is convex (concave) if the area under the curve is smaller (larger) than $1/2$. This calculation yields a characterisation of the individual utility function. For example, Qiu and Steiger (2011) classify subjects as having linear utility if the area is between 0.47 and 0.53, as having concave utility if the area is above 0.53, and as having convex utility if the area is below 0.47.

From Figure 7.15 it is also apparent that the slope of the segment lines between two successive elements of the standard sequence provides an approximation of the derivative of the utility function (i.e. marginal utility). As such, the 'rate of growth'

Focus 7.21
Comparing standard-gamble and value-equivalence methods

Wakker and Deneffe (1996), Bleichrodt et al. (2001) and Abdellaoui et al. (2007b) compare the utility elicited via the trade-off method to that from certainty equivalence. Wakker and Deneffe (1996) also measure utility via probability equivalence. Under expected utility, the results from the certainty-equivalence and trade-off methods should be consistent. Abdellaoui et al. (2007b) find no difference between the two methods when the probability associated with the best outcome is set to 1/3. This suggests that eliciting utility with a one-third probability helps to remove some of the usual departures from expected utility found in decision under risk (Bleichrodt et al., 2001). However, Abdellaoui et al. (2007b) also show that the results change dramatically under the certainty-equivalence method when the probability associated with the best outcome is 2/3. In this case, utility from the certainty-equivalence method differs from that in either the trade-off method or the certainty equivalence method with probability 1/3. As shown in the figure below (from Figure 3 in Abdellaoui et al., 2007b, p. 366), the average utility function from the certainty-equivalence method with probability 2/3, denoted $CE_{2/3}(EU$, is much more curved than that from the trade-off method (TO) or the certainty equivalent method with probability 1/3 ($CE_{1/3}$).



This finding is not in line with expected utility, and suggests that probability levels can have a considerable impact on elicitation. Abdellaoui et al. (2007b) reanalyse the results using prospect theory to take probability weighting in utility measurements into account. Under prospect theory, the resulting corrected utility, denoted $CE_{2/3}(PT)$, conforms with the other measurements (the figure also shows the results from eliciting utility from riskless strength-of-preference judgement, which is not discussed here). To conclude, standard-gamble methods, which involve a riskless prospect, are more prone to violations of

expected utility due to the well-documented certainty effect (Kahneman and Tversky, 1979; Fehr-Duda and Epper, 2012). This suggests that the utility elicited by these methods may be too risk-averse. In contrast, value-equivalence methods that are based on the comparison of risky prospects are less prone to certainty effects, and so might exhibit fewer violations of expected utility.

of the elements of the standard sequence directly characterises the shape of the utility function. Abdellaoui (2000) takes the first-order difference $\Delta_j^{'} = |x_j - x_{j-1}|, j = 1, \ldots 5$ and the second-order difference $\Delta_j^{''} = \Delta_{j+1}^{'} - \Delta_j^{'}, j = 1, \ldots 4$ to classify subjects by the shape of their utility function. The utility function is concave (convex) if and only if $\Delta_j^{''}$ is positive (negative). To account for response error, subjects with two out of four positive (negative) $\Delta_j^{''}$ are classified as having a concave (convex) utility function.

Last, the trade-off method yields parametric estimates of the utility function at low econometric cost. If we assume, for example, the power family, defined by (7.2), and a domain of utility normalised to [0, 1], estimation can be carried out using non-linear least squares in the following regression:[27]

$$\frac{j}{5} = \frac{x_j^{1-\theta}}{1-\theta} + \epsilon_j, j = 1 \ldots 5 \tag{7.12}$$

where, for example, $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$.

Most of the approaches presented in this section are based on the hypothesis that subjects behave according to expected utility. Under this assumption, utility measurement proceeds by eliciting a small number of equivalences between prospects. There is, however, a great deal of evidence that individuals systematically violate expected utility (Starmer, 2000) and that utility measurements based on expected utility produce inconsistent results. Two important sources of expected-utility violations are probability weighting – the non-linear evaluation of probabilities (Diecidue and Wakker, 2001) – and loss aversion, the assumption that people are more sensitive to losses than to commensurate gains. Loss aversion is key for the explanation of departures from expected utility (Rabin, 2000; Starmer, 2000) and has received abundant empirical evidence both from the lab and from the field (Wakker, 2010; Barberis, 2013; Fox and Poldrack, 2014). Both probability weighting and loss aversion are accounted for in prospect theory – developed in Focus 7.22. Focus 7.23 further details how loss aversion can be measured with the trade-off method.

Experimental Evidence

Wakker and Deneffe (1996) test the trade-off method for monetary and duration-of-life outcomes in two separate experiments. In the monetary experiment, participants are told that they could choose between two types of investment in a foreign country with an

uncertain return, and a five-element standard sequence is elicited. The results are consistent with risk aversion and utility that is concave in money. Qiu and Steiger (2011)

[27] With $\frac{(x_j)^{(1-\theta)}}{1-\theta}$ to be replaced by a logarithm if $\theta = 1$.

Focus 7.22
The basic prospect-theory model

For simplicity, we restrict the presentation to a two-outcome prospects model. Preferences are defined relative to a reference point (denoted x_0). Gains are outcomes that are strictly preferred to the reference point, while losses are outcomes that are strictly less preferred to the reference point. A prospect is mixed if it involves both a gain and a loss. For mixed prospects, the notation $x_1{}_p x_2$ stands for a situation in which x_1 is a gain and x_2 is a loss. A gain prospect involves no losses (i.e. both x_1 and x_2 are weakly preferred to x_0) and a loss prospect involves no gains. For gain and loss prospects the notation $x_1{}_p x_2$ applies to a situation in which the absolute value of x_1 exceeds the absolute value of x_2 (i.e. for gains $x_1 \geq x_2$ and for losses $x_1 \leq x_2$). Under prospect theory, the decision-maker's preferences over mixed prospects $x_1{}_p x_2$ are

$$U(x_1{}_p x_2) = \omega^+(p)v(x_1) + \omega^-(1-p))v(x_2) \tag{7.13}$$

and preferences over gain or loss prospects are

$$U(x_1{}_p x_2) = \omega^s(p)v(x) + [1 - \omega^s(p)]v(y) \tag{7.14}$$

where $s = +$ for gains and $s = -$ for losses, and v is a strictly increasing, real-valued utility function with $v(x_0) = 0$. In empirical applications, v is often decomposed into a basic utility function u, capturing the decision-maker's attitudes towards final outcomes, and a loss-aversion coefficient capturing attitudes towards gains and losses. The probability-weighting functions ω^+ and ω^- are strictly increasing and satisfy $\omega^+(0) = \omega^-(0) = 0$ and $\omega^+(1) = \omega^-(1) = 1$. Under prospect theory, risk attitudes are captured through the shape of the utility function for gains and losses, as in expected utility, but also through loss aversion (for mixed prospects) and probability weighting. As a consequence, the one-to-one relationship between risk aversion and utility curvature that exists under expected utility no longer holds under prospect theory. Tversky and Kahneman (1992) assume that the utility function and the probability-weighting functions ω^+ and ω^- show diminishing sensitivity. This produces an S-shaped utility function that is concave in gains and convex in losses, and an inverse S-shaped probability-weighting function that overweights small probabilities and underweights moderate and high probabilities (see Focus 5.2 and Focus 7.25). Taken together, the S-shaped utility and inverse S-shaped probability weighting imply a fourfold pattern of risk attitudes: risk aversion for small probability losses and larger probability gains, and risk seeking for larger probability losses and small probability gains.

replicate these findings in a larger experiment. For losses, Fennema and Van Assen (1998) and Etchart-Vincent (2004) find convex utility. Abdellaoui (2000) applies this method to gains and losses, and Bleichrodt and Pinto (2000) to the health domain. The results show concave utility for gains and convex utility for losses. These findings are consistent with a utility function exhibiting the psychological principle of diminishing sensitivity (Kahneman and Tversky, 1979). Booij and Van de Kuilen (2009) confirm these findings when measuring utility over gains and losses in a large representative sample of N = 1,935 Dutch respondents. Abdellaoui et al. (2016b) extend this method to capture sign-dependence in the utility function and elicit the utility function for gains

Focus 7.23
Measuring loss aversion

Measuring loss aversion is difficult for a number of reasons. First, there are a number of definitions in the theoretical literature – see Zank (2010) for a review and Abdellaoui et al. (2007a) for an empirical assessment of the various definitions. Second, measuring loss aversion requires the simultaneous measurement of utility for gains and utility for losses, which is complicated by the assumption of prospect theory that the probability weighting for gains and losses may differ. As a result, most measurements of loss aversion impose simplifying assumptions: these are typically linear utility for gains and losses and no probability weighting (Booij and Van de Kuilen, 2009; Baltussen et al., 2016), or equal probability weighting for gains and losses (Gächter et al., 2007). The typical coefficient of loss aversion found in the literature is around 2, meaning that losses weigh approximately twice as much as commensurate gains (see Fox and Poldrack, 2014, for a review of the available evidence). Abdellaoui et al. (2016b) introduce a method to measure loss aversion that can be applied under both risk and uncertainty, and requires no simplifying assumptions about the prospect-theory parameters. Moreover, the method allows for different definitions of loss aversion. In particular, under the definition of loss aversion proposed by Köbberling and Wakker (2005), the elicitation method can quantify loss aversion via three indifferences and does not require the complete measurement of utility. The method consists of three stages. In the first stage, a gain and a loss are elicited that connect utility for gains (measured in the second stage) with utility for losses (measured in the third stage). This first stage alone is enough if the aim is to measure loss aversion according to the definition in Köbberling and Wakker (2005), and works as follows.

- First, a reference point x_0 (e.g. 0) and a gain outcome x_G are set. The experiment then elicits the loss x_L for which the subject is indifferent between the reference point x_0 and a mixed prospect $x_G p x_L$, with a gain x_G with probability p and a loss x_L with probability $1-p$.
- Second, the experiment elicits two certainty equivalents: one in the gain domain x_1^+ and one in the loss domain x_1^-. The certainty equivalent in the gain domain is such that the subject is indifferent between receiving x_1^+ for sure and playing a lottery paying x_G with probability p and the reference point x_0 with probability $1-p$. The certainty equivalent for the loss domain is such that the subject is indifferent between losing x_1^- for sure and playing a lottery losing x_L with probability $1-p$ and paying the reference point with probability p.
- Abdellaoui et al. (2016b) show that under a fairly general specification of preferences, loss aversion as defined as the kink in utility at the reference point (Köbberling and Wakker, 2005) can be estimated directly by the ratio of the certainty equivalents x_1^+/x_1^-. If this ratio is greater than 1, the subject is loss-averse; if the ratio is below one, the subject is loss-seeking; and if the ratio is one, the subject is loss-neutral. Abdellaoui et al. (2016b) also show that replacing the probabilities p and $1-p$ by the (complementary) events with unknown probabilities allows loss aversion to be measured under uncertainty.

The second and third stages are necessary if the focus is rather on measuring loss aversion according to the definition by Kahneman and Tversky (1979). The measurements in these additional stages rely on Wakker and Deneffe's (1996) trade-off method explained in Section 7.4.5.

Focus 7.24
Prospect theory with uncertainty and ambiguity

Prospect theory, presented in Focus 7.22, can easily be extended to uncertainty and ambiguity (Tversky and Kahneman, 1992; Wakker, 2010). Uncertainty is modelled through a state space $\mathcal{S}$. Exactly one of the states will occur, but the decision-maker does not know which one. The subsets E of the state space are called events and E^c denotes the complement of E. Under uncertainty, prospects map states to outcomes: a two-outcome prospect is denoted $x_1 E x_2$, i.e. the decision-maker obtains x_1 if event E occurs and x_2 otherwise. Under binary prospect theory, the decision-maker's preferences over mixed prospects $x_1 E x_2$ are

$$U(x_1 E x_2) = W^+(E)v(x_1) + W^-(E^c)v(x_2) \tag{7.15}$$

and preferences over gain or loss prospects are

$$U(x_1 E x_2) = W^s(E)v(x_1) + [1 - W^s(E)]v(x_2) \tag{7.16}$$

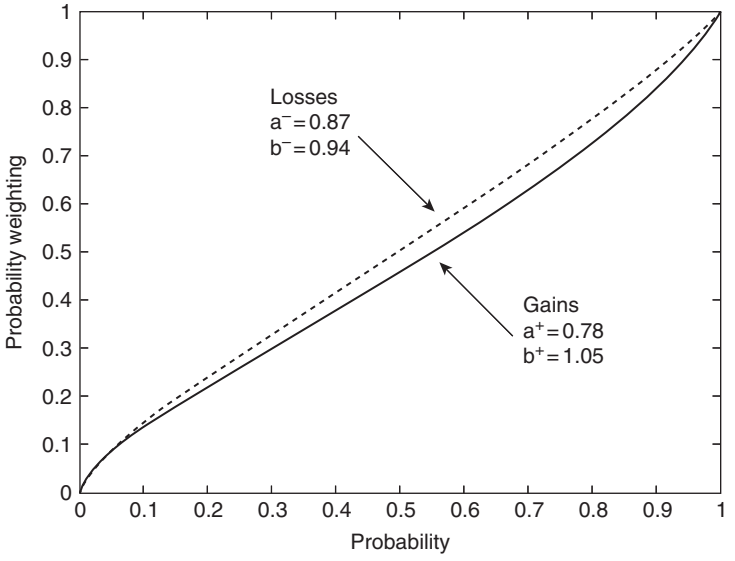
The event-weighting functions W^s assign a number $W^s(E)$ to each event E such that $W^s(\emptyset) = 0$, $W^s(\mathcal{S}) = 1$ and W^s is monotonic – i.e. $E \supseteq E'$ implies $W^s(E) \geq W^s(E')$. The event-weighting functions W^s depend on the sign of the outcomes and may be different for gains and losses. Moreover, they need not be additive. For gains, binary prospect theory contains most transitive ambiguity models as special cases, as pointed out by Luce (2014). These ambiguity models only differ when the number of outcomes is at least three; (7.15) and (7.16) provide the extension of these models to include sign dependence. For binary prospects, Baillon et al. (2017) show how to describe a decision-maker's ambiguity attitude by two indices to measure and decompose ambiguity attitudes in experiments.

and losses in a within-subject experiment under risk and uncertainty. Van Assen and Snijders (2010) elicit subjects' utilities by the trade-off method and use these utilities to make predictions in an experiment in which participants played a number of different repeated prisoners' dilemma games. They find no evidence that concave utility produces cooperation in the prisoners' dilemma.

The trade-off method is not restricted to choice under risk, and can be easily extended to choice under uncertainty. For decision under uncertainty, the trade-off method is similar to that above with the probability p being replaced by an event with unknown probability E. For example, the indifference in (7.7) becomes $x_1 E 1 \sim 10_E 8$, meaning that the subject is indifferent between receiving x_1 if event E occurs and 1 otherwise; and receiving 10 if event E occurs and 8 otherwise. In this case, under (subjective) expected utility, the objective probabilities have to be replaced by subjective probabilities $Pr[E]$ and $1 - Pr[E]$. As the probabilities disappear when (7.9) and (7.10) are combined, the trade-off sequence defined by (7.11) remains the same. Wakker and Deneffe (1996) use electoral outcomes to describe the event E and its complement, and Abdellaoui et al. (2005) use movements in the German stock market. Abdellaoui et al. (2005) obtain concave utility for gains and close to linear utility for losses. Abdellaoui et al. (2016b) elicit the utility function for the same individuals under both risk and uncertainty on a within-subject basis. In both contexts, utility is mostly concave for gains and convex for losses.

Focus 7.25
Probability weighting in choice under risk

Most empirical work on probability weighting has found an inverse S-shaped probability weighting for both gains and losses (Tversky and Kahneman, 1992; Wu and Gonzalez, 1996; Abdellaoui, 2000; Bleichrodt and Pinto, 2000; Booij et al., 2010; Tanaka et al., 2010; Bruhin et al., 2010; Abdellaoui et al., 2011). The following figure (from data on US students in L'Haridon and Vieider, 2015) is a typical example of probability-weighting functions for gains and losses, obtained by fitting the two-parameter weighting function proposed by Prelec (1998), $\omega^s(p) = e^{-b^s(-ln(p))^{a^s}}$.



The parameter a mainly determines the slope of the probability-weighting function: $a = 1$ if the weighting function is linear (the EU case) and $a < 1$ in the case of probabilistic insensitivity. Any value of a that differs from 1 indicates a departure from linear utility weighting and hence from EU. The parameter b is an anti-index of the elevation of the weighting function, i.e. the extent to which people are optimistic (or pessimistic) and find the chance domain attractive. The point where the probability-weighting function changes from overweighting probabilities to underweighting them is typically around $^1/_3$.

For losses, the equal curvature of utility for risk and uncertainty cannot be rejected; for gains, utility is significantly more concave under risk.

7.4.6 The Econometrics of Individual Preferences under Uncertainty

The econometric procedures discussed above have been widely used in experimental economics to estimate risk-preference parameters (Holt and Sherman, 2014; Hey, 2014; Moffatt, 2015). Several extensions have been proposed to infer preference parameters

from binary choices or indifferences between lotteries. In this section, we review some of these extensions, and focus on two specific dimensions: error specification and heterogeneity in risky choice.

Error Specification and Randomness

In Loomes et al. (2002) the choice between two prospects is analysed in three stages: preference selection, calculation and action. Randomness can enter the choice between prospects at any of these three stages. At the first stage, there is randomness if the subject is uncertain about their preferences. In the random-preference model of Loomes and Sugden (1995), preferences are defined according to a core theory (expected utility or prospect theory, for example) but the parameters applied to a given choice vary randomly. There is randomness at the second stage if the decision-maker makes calculation errors when comparing prospects: this corresponds to the standard representation of stochastic choice. Last, at the third stage, the decision-maker can fail to choose the prospect that they thought had the highest value – because, for instance, of a lack of focus when making decisions, or choices that are made at random, or because decision-makers did not understand the decision problem. In this case, the decision-maker's hand trembles and the choice becomes less strongly connected with preferences. We start by extending the representation proposed in Section 7.4.4 to trembling hands (the third stage in Loomes et al., 2002) and then suggest different specifications for the stochastic term (the second stage in Loomes et al., 2002). Focus 7.26 contains more details on stochastic choice.

Introducing trembles allows mistakes to be taken into account once the calculation stage has been performed (inattention, true random choice, violation of first-order stochastic dominance, etc.) when choosing between prospects. In Moffatt and Peters (2001), for example, there is a probability p_τ that subjects choose at random between lotteries (with probability $1/2$ in the case of a binary choice) while subjects choose according to the standard choice probability $Pr[A_j]$ with probability $1-p_\tau$. The choice probability then becomes

$$Pr[A_j|p_\tau] = (1-p_\tau)Pr[A_j] + \frac{p_\tau}{2} \tag{7.19}$$

If the choice probability $Pr[A_j]$ depends on some preference parameter θ and a stochastic term σ, then the base choice probabilities is $Pr[A_j|\theta,\sigma]$ and the choice probability becomes

$$Pr[A_j|\theta,\sigma,p_\tau)] = (1-p_\tau)Pr[A_j|\theta,\sigma] + \frac{p_\tau}{2} \tag{7.20}$$

Using data from the experiment in Hey and Orme (1994), Moffatt and Peters (2001) evaluate the tremble probability p_τ to be around 3%, a value halfway between 1% in Conte et al. (2011) and 8% in Von Gaudecker et al. (2011). Loomes et al. (2002) find that trembling falls with experience: the tremble is 11% at the beginning of the experiment but under 2% at the end.

Different specifications of calculation errors have been used in the literature. The basic specification (Fechner, 1869) assumes it to be independently and identically distributed, and normally distributed with constant variance. Denoting this calculation error

Focus 7.26
Stochastic choice

As shown in Section 7.4.4 it is necessary to add a stochastic component ϵ to the analysis if some randomness is embedded in choices. In this framework, the decision-maker chooses A_j over B_j if and only if $U(A_j) - U(B_j) + \epsilon_j > 0$. The standard interpretation is that ϵ captures random unobservable characteristics (Train, 2009). The model can be further specified by denoting $g(.)$ the density function, and $G(.)$ the cumulative distribution function, of the random variable ϵ_j. Let $Z_j = 1$ when the value of ϵ_j together with the value of observables lead the decision-maker to choose A_j, and 0 otherwise. Making use of the assumed decision process, the probability of choosing A_j is the expected value of the indicator over the set of possible values of ϵ_j:

$$\begin{aligned} Pr[A_j] &= \int Z_j[U(A_j) - U(B_j) + \epsilon_j > 0]g(\epsilon_j)d\epsilon_j \\ &= \int Z_j[\epsilon_j > -[U(A_j) - U(B_j)]]g(\epsilon_j)d\epsilon_j \end{aligned}$$

As $Z_j = 0$ when $\epsilon_j < -[U(A_j) - U(B_j)]$, the integral only takes into account observations such that $Z_j = 1$ (i.e. when $\epsilon > -[U(A_j) - U(B_j)]$, and the probability is

$$Pr[A_j] = \int_{-[U(A_j)-U(B_j)]}^{\infty} g(\epsilon_j)d\epsilon_j \tag{7.17}$$

Then

$$Pr[A_j] = 1 - G[-[U(A_j) - U(B_j)]] \tag{7.18}$$

If a normal distribution is assumed for the error, symmetry implies $Pr[A_j] = G[U(A_j)-U(B_j)]$. With μ_j and σ_j as the parameters of the normal distribution ($\epsilon_j \sim N(0, \sigma_j)$), then $Pr[A_j] = \Phi\left(\frac{U(A_j)-U(B_j)}{\sigma_j}\right)$.

ϵ_j, this corresponds to assuming $\epsilon_j \sim \mathcal{N}(0, \sigma^2), \forall j$. As such, the probability of selecting A_j over B_j is

$$Pr[A_j|\theta, \sigma] = \Phi\left[\frac{U(A_j|\theta) - U(B_j|\theta)}{\sigma}\right] \tag{7.21}$$

where Φ denotes the unit normal cumulative distribution function and θ preference parameter(s). In Section 7.4.4, the preference parameter was the CRRA coefficient θ. As the variance of the error parameter σ^2 tends to 0 (no error), choice becomes deterministic and $Pr[A_j|\theta, \sigma]$ tends to 1 or 0 depending on the sign of $U(A_j|\theta) - U(B_j|\theta)$. With high variance, on the contrary, subjects make their choices essentially at random and $Pr[A_j|\theta, \sigma]$ tends to one-half.

Von Gaudecker et al. (2011) define the choice probabilities (7.21) based on certainty equivalence between lotteries. For a utility function $u(.)$ and utility parameter θ, the certainty equivalent of option A_j is $CE(A_j|\theta) = u^{-1}\left[U\left(A_j|\theta\right)|\theta\right]$ and that of option B_j is $CE(B_j|\theta) = u^{-1}[U(B_j|\theta)|\theta]$. In this case, the variance σ has a monetary interpretation and the choice probabilities are

$$Pr[A_j|\theta,\sigma] = \Phi\left[\frac{CE(A_j|\theta) - CE(B_j|\theta)}{\sigma}\right] \tag{7.22}$$

When choices have been observed between a fixed risky prospect A_j and a sequence of certain alternatives B_j, the switching point between the two options defines the certainty equivalent of the lottery. The choice probabilities can be defined in two ways. Bruhin et al. (2010) take the middle of the switching interval as the (observed) certainty equivalent CE_j. The probability that the predicted certainty equivalent $CE(A_j|\theta)$ matches the elicited certainty equivalent is

$$Pr[CE(A_j) = CE_j|\theta,\sigma] = \frac{1}{\sigma}\phi\left(\frac{CE_j - CE(A_j|\theta)}{\sigma}\right)$$

where ϕ denotes the standard normal density. Beauchamp et al. (2012) use the two certain options B_{inf} and B_{sup} that bracket the certainty equivalent to define the choice probability. The probability that the participant switches between $B_{j,inf}$ and $B_{j,sup}$ is

$$\begin{aligned} Pr[B_{j,inf}, B_{j,sup}|\theta,\sigma] = \; & \Phi\left(\tfrac{U(B_{j,sup}|\theta) - U(A_j|\theta)}{\sigma}\right) \\ & -\Phi\left(\tfrac{U(B_{j,inf} - U(A_j|\theta)|\theta)}{\sigma}\right) \end{aligned}$$

An extreme-value distribution, such that $\epsilon_j \sim \text{EV}(0,\sigma^2)$, can be assumed instead of *i.i.d.* normally distributed errors. This error specification leads to a mixed-logit model and proves to be very useful in computationally intensive estimations (Train, 2009). The probability of selecting A_j for choice j in this case is

$$Pr[A_j|\theta,\sigma] = \frac{1}{1 + e^{-\frac{U(A_j|\theta) - U(B_j|\theta)}{\sigma}}} \tag{7.23}$$

There is a link between the extreme-value distribution of error and the Luce error model (Wilcox, 2008). Applying a logarithmic transformation of $U(.)$, $V(.|\theta) = log[U(.|\theta)]$ results in the Luce error model, with $\mu = \sigma$:

$$Pr[A_j|\theta,\sigma] = \frac{V(A_j|\theta)^{\frac{1}{\sigma}}}{V(A_j|\theta)^{\frac{1}{\sigma}} + V(A_j|\theta)^{\frac{1}{\sigma}}}$$

All the above specifications assume homoskedastic variance. The error parameter σ can then be decomposed into a component common to all choices and a measure of distance between prospects. Wilcox (2008) suggests two alternative ways of including heteroskedasticity into the error specification. A first is to link the error parameter to a distance between outcomes. For example, in choices between a risky prospect $x_{1j_p}x_{2j}$ and certain amounts, Bruhin et al. (2010) make the error parameter proportional to the range $|x_{1j} - x_{2j}|$ between outcomes: $\sigma_j = \sigma \times |x_{1j} - x_{2j}|$. Another possibility is to link the error parameter to a distance between the utilities of outcomes.

Instead of assuming identically distributed variance between successive tasks, a task-varying error component can be specified. This assumption can reflect learning or experience (if the error variance falls with experimental duration) or fatigue (if error variance rises with experimental duration). Assume that the experimental choices ($j = 1,\ldots,J$) are indexed by the order in which the subject performs the decision tasks.

One possible specification for a varying error component is exponential: $\sigma_j = \sigma_0 \times e^{\sigma_1 j}$, where the parameter σ_0 indicates the error parameter at the beginning of the experiment and σ_1 shows how the error changes throughout the experiment. If $\sigma_1 > 0$, then the subject tires as the experiment progresses. If $\sigma_1 < 0$, then the subject learns and makes fewer errors.

Heterogeneity

Starting from Hey and Orme (1994) and Harless and Camerer (1994), a series of papers use binary choices or indifferences between lotteries to consider heterogeneous risk preferences. A first way is to estimate preferences separately for each individual for a given decision model (e.g. expected utility). An extension is to estimate the preference parameters for different decision models and select, for each subject, the best-fitting model using a likelihood-ratio test or the Akaike information criterion. Hey and Orme (1994) consider a series of 100 binary choices by subjects to identify the best-fitting individual preferences by maximum likelihood. Stott (2006) uses the same method and evaluates a pattern of 256 combinations of decision models and stochastic choice patterns to find the best-fitting combination over the 96 experimental participants.

A second way of dealing with heterogeneity is to assume that preferences are fixed for a reference individual and vary with some observable characteristics. Here the preferences θ_i of subject i are conditional on a set of observable characteristics, X_i:

$$\theta_i = \boldsymbol{X}_i \mathbf{b} \tag{7.24}$$

where $\boldsymbol{X}_i$ is a vector of regressors with the first element being one, and $\mathbf{b}$ is a parameter vector. Harrison and Rutström (2008) use this specification to control for individual characteristics in the measurement of the risk parameter θ using the Holt and Laury (2002) method.

A third way to deal with heterogeneity is to estimate a random-coefficients model in which the preference parameter is fixed for a given individual and varies randomly across the population according to a given distribution (Loomes et al., 2002). For example, Moffatt and Peters (2001) assume that the power/CRRA parameter $1 - \theta$ is lognormally distributed across the population, and Conte et al. (2011) assume that the parameters from a two-parameter rank-dependent utility model have a bivariate distribution. If θ is drawn from a distribution with density $g(\theta)$, then for a given set of decisions $j = 1, \ldots, J$, the likelihood associated with subject i's choices, Z_{ij}, is

$$L_i(\theta, \sigma) = \int \left[\prod_{j=1}^{J} Pr[A_j|\theta, \sigma]^{Z_{ij}} \times (1 - Pr[A_j|\theta, \sigma])^{1-Z_{ij}} g(\theta) \right] d\theta \tag{7.25}$$

and the overall log likelihood is the sum of the logarithm of $L_i(\theta, \sigma)$ over all subjects. For the sake of notational simplicity, the stochastic parameter σ in (7.25) is assumed to be constant. If there is more than one preference parameter, as is often the case in non-expected-utility models, then the distribution $g(.)$ is a multivariate distribution and the integral in (7.25) is multidimensional. Von Gaudecker et al. (2011) estimate a random-coefficients model with four preference parameters. Each parameter is assumed to be

the (transformed) sum of a term reflecting observed heterogeneity in the spirit of (7.24) and a normally distributed unobserved heterogeneity term that is independent of the regressors.[28] In general, the integral in (7.25) does not have an analytical solution and is approximated by simulation techniques (e.g. Monte Carlo, Halton draws). Simulation methods approximate non-linear estimation problems by generating random draws that aim to cover the area of integration. A typical simulation method for estimating, e.g. a model with parameters distributed according to the density $g()$ of a normal distribution $\mathcal{N}(\mu, \psi)$, proceeds as follows (Train, 2009; Moffatt, 2015):

1. Take a draw of θ from $\mathcal{N}(\mu, \psi)$. Label this draw θ_1, where the subscript denotes that it is the first draw.
2. Determine $Pr[A_j|\theta, \sigma]$ with this value of θ_1 and for the value of σ and calculate $l_{i1} = \prod_{j=1}^{J} Pr[A_j|\theta_1, \sigma]^{Z_{ij}} \times (1 - Pr[A_j|\theta_1, \sigma])^{1-Z_{ij}}$ for each subject i over the J choices.
3. Repeat steps 1 and 2 many times, for a total of R draws, and compute the corresponding $l_{ir}, r = 1....R$.
4. Compute the average of the l_{ir}s, , $r = 1....R$. This average is the simulated individual contribution to the likelihood, $L_i(\theta, \sigma)$. Sum the log of this quantity over subjects to obtain the overall log likelihood for a given μ, ψ and σ.
5. Iterate over μ, ψ and σ using standard algorithms (BFGS, for example) to obtain the maximum likelihood.

A fourth way of dealing with heterogeneity is to use a finite-mixture model. This classifies individuals according to a given number of types in the population. Each type $h = 1, \ldots, H$ is characterised by a specific preference parameter θ_h and each subject has probability p_h of belonging to type h.[29] For a given type h and a given set of decisions $j = 1, \ldots, J$, the likelihood associated with subject i's choices Z_{ij} is

$$L_i(\theta_h, \sigma) = \prod_{j=1}^{J} Pr[A_j|\theta_h, \sigma]^{Z_{ij}} \times \left[1 - Pr\left[A_j|\theta_h, \sigma\right]\right]^{1-Z_{ij}} \tag{7.26}$$

and the overall log likelihood of the finite-mixture regression model is

$$LL(\theta_1, \ldots, \theta_h, p_1, \ldots, p_{h-1}, \sigma) = \sum_{i=1}^{N} \sum_{h=1}^{H} p_h L_i(\theta_h, \sigma) \tag{7.27}$$

For the sake of notational simplicity, (7.26) and (7.27) assume that the stochastic parameter σ is constant. On the contrary, Bruhin et al. (2010) assume that the error parameter is heteroskedastic and subject-dependent. They estimate the model using an iterative expectation maximisation (EM) algorithm for a different number of types. They evaluate the quality of classification via the average normalised entropy criterion and find that $H = 2$ provides a satisfactory account of the heterogeneity in their data. Bruhin et al.

[28] The sum of observed and unobserved heterogeneity is transformed to take restrictions on the sign or values of the parameters into account.

[29] It is also possible to define the type by assuming that different individuals use different preference functions (Bruhin et al., 2010; Conte et al., 2011).

(2010) and Conte et al. (2011) use finite-mixture models to characterise behaviour under two alternative decision models: expected utility and prospect theory. Each experiment has a different experimental design. Using a certainty-equivalence method, Bruhin et al. (2010) find that 80% of subjects can be classified as prospect-theory type with an inverse S-shaped probability weighting function. Using a paired-gamble comparison method, in which subjects are known to be less prone to violations of expected utility, Conte et al. (2011) find the opposite proportions.

Summary

This chapter provided an overview of statistical and econometric techniques that are commonly used with experimental data. Starting from a working example of second-price auctions, we first described what experimental data are in practice, depending on their data types and the way sampling was achieved. The chapter paid specific attention to exploratory data analysis, which is a fundamental way to get basic information from the experimental data. Graphical tools, in conjunction with descriptive statistics, are central for the researcher to learn the content of experimental data. In analysing experimental data, measures of association, such as the Pearson correlation, should be used with care. Anscombe's quartet clearly shows the importance of underlying assumptions in measurement and in the choice of the relevant indicator.

The chapter then moved to inference (as defined and discussed in Chapter 3) from a practical point of view. Several cases for building confidence intervals for various estimates, alongside the properties of estimators, are described. An important part of the chapter is devoted to hypothesis testing, a central analytical tool in experimental research. Hypothesis testing allows us to make a judgement about a parameter value by setting this judgement up as a testing problem between two competing, mutually exclusive, hypotheses regarding the true parameter value. Because experimental samples are often of limited size, we draw particular attention to non-parametric hypothesis tests and to issues related to sample size. This chapter does not intend to be exhaustive, but rather to account for the main statistical tests available in the literature, depending on the nature of the data (interval, nominal or categorical for example, or paired versus non-paired data) and depending on the objective of the test. The last section is devoted to the elicitation of risk preferences. This case study provides an overview of the experimental data that are often encountered in the measurement of risk preferences. Based on various decision theories, it also illustrates the main econometric and statistical techniques that can be applied to experimental data in order to elicit or to estimate risk preference parameters.

Part IV

What For? What Laboratory Experiments Tell Us

8 The External Validity of Experimental Results

The first question that comes to mind when wondering about how useful experiments are is: what about real life? This question is methodologically sensitive because of the specification of the design, chosen to enhance internal validity. The design of the experiment is the experimental data-generating process. Internal validity requires pursuing two (opposite) goals: identification and precision. Both goals require knowledge about the noise in the empirical measure. The first is achieved by breaking any correlation between the noise component and the experimental treatment and other control variables required for identification purposes. The second requires that we hold the maximum of variables constant, in order to maximise precision. As discussed at length in Chapter 5, these concerns lead to the design of experimental situations that are as abstract, and to some extent as far removed from their natural circumstances, as possible. This reflects that most dimensions of social phenomena appear as noise with respect to the single dimension under investigation. A key requirement for clean empirical measurement in the laboratory is the proposition of a situation that is as different as possible from the real-world situation of interest – this is why the laboratory is the appropriate place for the empirical investigation of the mechanisms behind real-world phenomena, as shown in Illustration 8.1.

External validity refers to the naturally occurring question once such measurement has been carried out: what is the point of establishing accurate measurement if it has nothing to do with social behaviour in real life? This chapter addresses this difficult question in a number of steps. Section 8.1 shows that this (apparently) simple question cannot easily be converted into a single definition of external validity. The question rather leads to a number of different definitions, or degrees, of external validity, each depending on the nature of the research question. Section 8.2 turns to the question of how external validity can be assessed. Whatever the definition, external validity boils down to an empirical question, as it requires the comparison of two empirical situations (one in the laboratory, the other in the 'real world'). The choice of the best empirical method to ascertain external validity is thus the main question to be addressed. Section 8.2 describes the main features of laboratory experiments that may challenge external validity, and reviews how these have been addressed in empirical research. Empirical tests of external validity always involve the behavioural consequences of some change in the design of the empirical analysis. As such, concerns about the external validity of experimental results call for pledges for more replication studies, which is the focus of the case study we present in this chapter.

Illustration 8.1

Reversed external validity: experimental evidence on the winner's curse in real auction markets

Auctions are called 'common value' when all bidders on the market have the same valuation of the good, as opposed to private-value auctions where the values are bidder-specific. If the common value is perfectly known, the outcome is trivial as bidders simply bid this common value and the market clears based on the tie rule. The literature has thus focused on the special case in which the bidders only imperfectly know the common value. Typically, each bidder receives a signal about the value of the good that is partially informative. Concretely, consider a two-bidder market ($i = 1, 2$), each with a signal ε_i uniformly drawn on $[0, 1]$, and for which the true value is $\varepsilon^* = (\varepsilon_1 + \varepsilon_2)/2$. Bidders suffer from a winner's curse in this market: the bidder with the highest signal will bid more than the true value of the good. This risk will be accommodated in equilibrium by bids that are lower than the signal received. This set-up has long been recognised as relevant in many market settings. For example, Capen et al. (1971) analyse oil-lease auctions by oil companies operating in the Gulf of Mexico. They show that the industry rate of return is consistent with the lease winner having the greatest overestimate of the resource. This is puzzling from the point of view of game theory. Experienced bidders with considerable monetary incentives, such as oil companies, should have learned about the winner's curse. To investigate, Kagel and Levin (1986) designed a first-price, sealed-bid, common-value experimental auction. Players privately submit their bid, and the winner pays the price of their bid. Players were found to bid most of the time over the optimal strategy. In all but one experiment, average profits were lower than the Nash prediction, and 34.8% of each experiment's bids exceeded the expectation of the object's value conditional on winning. Bidders were, however, found to learn from repeated exposure to the same market conditions, with bids that were closer to the Nash equilibrium after several rounds of play. Moreover, on markets with fewer than five bidders, profits were positive and on average equal to 65.1%. Learning was, however, limited to this context: the observed bids became even more aggressive after learning with higher numbers of bidders, contrary to the Nash equilibrium strategy. With over five bidders, average profits became negative. Kagel and Levin (1986) suggest that bidders are able to avoid the winner's curse in small groups, but do not show a real understanding of the winner's curse problem.

8.1 When and How Does External Validity Matter?

The question of external validity is easily seen to be both important and relevant: to what extent can the results of a study be generalised to other situations and other individuals? An early statement was that of Campbell and Stanley (1963, p. 5), who define external validity as the question of 'to what populations, settings, treatment variables, and measurement variables can this effect be generalised?' Manski (1999, p. 17) provides a more statistically oriented definition, which clearly reveals its underlying complexity:

'an invariance assumption, used to apply outcomes of social experiments to predict the outcomes of actual social programs. Distribution of outcomes realised by the treatment group is the same as the distribution of outcomes that would be realised in an actual program'.

When assessing external validity, 'external' is the key word, first because the question is always whether the result observed inside the laboratory continues to hold 'outside' it and second, because this underlies the difference between the two core validity issues: internal (as discussed in Chapter 5) and external validity. In a nutshell, internal validity is related to inference and identification inside the laboratory, and external validity concerns inference and identification outside the laboratory. The requirement for external validity is not self-evident. After all, why would we want to generalise observations that were made in a well-defined and controlled environment? The desire to move outside the laboratory reflects the need to understand what experiments tell us about the real world. Unfortunately, while the lab environment is easily controlled, the same does not hold for the 'real world'. The difficulty here stems from there being as many definitions of external validity as there are steps from the laboratory to the real-world counterpart. This section describes the range of such definitions and identifies the associated challenges regarding the ability of laboratory experiments to inform us about real-world situations.

8.1.1 The Many Meanings of External Validity

The existing definitions of external validity range from weak to far stronger requirements. As we will see, the assessment of the external validity of an experiment requires this kind of plurality of definitions, as the choice of definition is closely related to the inference that we would like to make.[1]

To help organise the discussion of external validity, it is useful to return to the definition of an experiment in Chapter 4. An experiment is a pseudo-real situation made up of two different types of input: the controlled (i.e. chosen) inputs, denoted by x^n, and all of the others, $x^{\infty-n}$. Among the infinite possible consequences, the experiment defines a subset of consequences of interest that are actually measured. An experiment can then be summarised by the functional $F^m(x^n \cup x^{\infty-n})$, mapping a set of inputs to its observed consequences. External validity refers to the relationship of the experiment to its real-world counterpart. We define the real world as another functional G, resulting from an infinite number of inputs (standing for the determinants of a given situation) producing an infinite number of consequences $G^{\infty}(X^n \cup X^{\infty-n})$. All of the definitions of external validity compare the two functions F and G, to assess whether laboratory experiments

[1] This question has recently became very controversial and main field of battle between those for and against laboratory experiments in economics. It is, in addition, a deep epistemological question, referring to no less than the scientific contribution of empirical evidence. For this perspective on external validity, we refer the reader to specialised contributions such as Guala (2005).

Focus 8.1

The many different meanings of external validity in experimental psychology

There is one exception to the many differences in experimental methodology between economics and psychology: as in economics, the concept of external validity in psychology has many different definitions and is still the subject of vivid debate. Perhaps the main difference between the two fields in this context is that external validity has been discussed for longer in psychology, and remains the subject of many methodological contributions. The most common notion of external validity in psychology is 'ecological validity'. This use of the terminology is, however, significantly removed from that in the original definition in Brunswick (1956), who defined ecological validity as the informational value of a signal received by a decision-maker (see Hammond, 1998, for a refreshing and insightful historical perspective). Currently, 'ecological validity' often refers to the real-world conditions of the phenomenon studied in the experiment (see e.g. Schmuckler, 2001, for an example in cognitive psychology). This definition is in fact closer to the terminology in Brunswick (1956) of a 'representative design', which refers to an experimental design regarding a well-defined situation, where the components of the experiment are representative. Design representativeness, referred to as 'ecological validity', is, for instance, part of the two main notions of external validity in psychology in the survey by Brewer and Crano (2014) along with 'robustness' – the ability to replicate a result in different conditions or with different samples. In addition to representativeness and robustness, a third notion, 'relevance', considers whether a finding actually applies to real-world problems.

inform us about real-world situations. The many different meanings of validity differ in how components are matched between the two. As shown in Focus 8.1, this plurality of definitions is not specific to economics, and is shared in particular with experimental psychology. Parallelism is the most restrictive definition, asking whether experimental results will continue to hold in the real world. Two other definitions refer more clearly to generalisability, which aims to establish the range of input values for which the laboratory results remain valid outside the laboratory. These definitions produce a grey zone, with which we will conclude this section, regarding the link between external validity and the target parameter of interest.

Definition 1: parallelism.

One of the first formal discussions of external validity in the experimental-economics literature referred to a restrictive definition known as *parallelism* (Smith, 1982, p. 936). Parallelism restricts external validity to the question whether 'propositions about the behaviour of individuals and the performance of institutions that have been tested in laboratory micro-economies apply also to non-laboratory micro-economies where similar *ceteris paribus* conditions hold'. The question can be rephrased as whether F^m coincides with $G^m \in G^{\infty}(X^n \cup X^{\infty-n})$ when $x^n \equiv X^n$. The two situations are parallel, according to this particular definition of external validity, if the causal relationship identified in the laboratory continues to hold outside it under exactly the same circumstances – i.e. when the controlled inputs are set to the same values in the real world as inside the

laboratory. In the example of gift exchange in work situations discussed in Chapter 4, Section 4.5.2, the field experiment of Gneezy and List (2006) does exhibit this weak form of external validity, as the same phenomenon occurs in the real-world environment that strictly replicates the laboratory setting – i.e. during the first 90 minutes of the field experiment, which corresponds to the maximum typical duration of the laboratory experiment. This form of external validity is also substantiated by the field experiment in Falk (2007), finding higher donations when a gift is enclosed with solicitation letters, and hence substantiating reciprocal behaviour in one-shot relationships in real life.

The main challenge faced by laboratory experiments with this weak form of external validity is what Schram (2005) calls their *artificiality*. The desire to have as much control as possible over the environment in order to help ensure internal validity leads to the jettisoning of as many features of the real world as possible.[2] The use of neutral instructions with no reference to the actual context is the most obvious example of this artificiality – raising the external-validity concern that behaviour in this abstract environment might not be replicated when the same inputs are returned to their 'natural' surroundings. In other words, does the laboratory environment produce behaviour that would not occur under the same conditions outside the laboratory? Some of the criticisms of the external validity of the Zimbardo 'Stanford prison' experiment discussed in Chapter 5, Illustration 5.9, are along these lines. Many of the students who took part in the experiment explained that they were only acting consistently with their role of 'guard' or 'prisoner'. In this reading of the data, the assignment of roles and the need to underline their implications in order to explain the experiment to subjects drove behaviour that would not have occurred under natural conditions. Similarly, Illustration 8.2 describes an experimental test of the external validity of a very artificial set of experiments, testing corruption behaviour based on modified gift-exchange games.

Definition 2: robustness.

The artificiality of the experimental environment also raises a closely related, but very different, series of questions. The inputs are chosen following the requirements of the experimental environment, but the experiment itself aims to replicate some relevant real-world situation. The gift-exchange game, for instance, aims to import to the laboratory some features of an employment situation: the interaction between an employer and an employee, their different pecuniary interests, their joint production of economic surplus, and the way in which the surplus is split between the two parties. The laboratory implementation reduces the real-world situation, in both scope – as many inputs are neglected by design – and size – due to physical and/or methodological constraints. The question of external validity here thus requires us to assess whether this restrictive definition of the inputs affects the observed outcomes.

Robustness in applied work refers to the replication of an observed phenomenon in an alternative context or a different empirical framework using similar values as inputs. The causal mechanism of interest remains the same, and the empirical question is

[2] Identification requires us to construct experiments with highly abstract and artificial institutions, which resemble their real-world counterparts less and less as greater control introduces a setting that becomes increasingly lab-specific.

Illustration 8.2
The measure of corruption from laboratory bribery behaviour

Laboratory experiments are a natural way of collecting data on antisocial or illegal behaviour, given the obvious limitations to their observation in the field. Among the possible challenges to the external validity of corruption experiments, the most common are the failure to account for non-monetary motives (e.g. ethical and legal motives), the effect of subjects' awareness of being monitored, and the fact that some relevant real-life features may not be reproducible in a laboratory. Armantier and Boly (2013) aim to test the relevance of these factors using three different versions of a corruption experiment: a laboratory experiment in a developed country, a laboratory experiment in a developing country and a field experiment in this same developing country. The first design step is to define a task allowing the observation of corruption that can easily be replicated both in the laboratory and in the field. To this end, the experiment considers the behaviour of a grader who is offered a bribe to over-grade an exercise paper. As a preliminary stage, some individuals were recruited in Montreal to take part in a dictation. Seven papers, with varying numbers of mistakes, were selected and completed with 13 artificial papers, made up by the experimenters: this set of 20 papers constituted the pool of papers to be graded by the participants. In the field experiment, which took place in Ouagadougou, Burkina Faso, people were recruited via flyers, offering a part-time job for university students. Subjects were not aware that they were taking part in an experiment. They were asked to spellcheck the 20 papers, to report the number of mistakes in each of them and to decide on the final result, pass or fail, with a passing threshold fixed at 15 mistakes. The grading took place individually, in a closed room. The 11th paper from the stack came with some money and a sticky note: 'Please, find only a few mistakes in my examination paper'. Graders who informed the supervisor of the bribe were asked to grade the paper as a fail. Once the grading was over, participants were told that they were in an experiment and were paid the announced rate. In addition to the field experiment, two laboratory experiments took place in Ouagadougou and in Montreal, Canada, based on exactly the same protocol, with the exception that people knew in advance that they were in an experiment. In each of these three experiments, subjects were randomly assigned to one of four treatments: CONTROL, HIGH BRIBE, HIGH WAGE and MONITORING, in which subjects face a potential double-check of some of the graded papers by the supervisor. The main results appear in the following table (from Armantier and Boly, 2013, p. 10, Table 2); stars indicate significant treatment effects ($*$, significant at the 10% level, $**$, significant at the 5% level), and daggers a significant difference from the same treatment in the laboratory in Ouagadougou (†, significant at the 10% level, ‡, significant at the 5% level).

The outcomes obtained in the laboratory and field in Ouagadougou are very similar. Scrutiny does not seem to have any distortionary effect. The reactions to the treatment variables are also in line with what was expected, and of comparable size across environments and countries, suggesting that results from a developed country can be extrapolated to a developing country. In particular, higher wages unambiguously reduce corruption, but promote reciprocity (i.e. over-grading of the 11th paper) by the bribe-takers.

Environment	Treatment	No. of Subjects	% of graders who accept the bribe	Average no. of mistakes reported for paper 11	
				Accepters	Rejecters
Field (Ouagadougou)	Control	36	50.0	14.7	16.4
	High bribe	45	68.9**	13.1‡	15.4
	High wage	39	35.9*	12.9†	15.7‡
	Monitoring	44	40.9	15.7†	15.9
Lab (Ouagadougou)	Control	33	48.5	16.4	17.3
	High bribe	33	66.7**	15.2	16.6
	High wage	33	36.4*	14.9	17.5
	Monitoring	34	41.4	17.5	16.5
Lab (Montreal)	Control	30	66.7†	14.9	16.7
	High Bribe	32	65.6	14.8	15.6
	High Wage	31	48.4**	13.9	16.3†
	Monitoring	32	65.6‡	15.0‡	15.5

the extent to which the occurrence of this mechanism depends on the context in which it is investigated. In econometrics, for instance, robustness checks include changing the set of control variables, applying alternative estimation models or replicating the analysis on data sets from alternative sources. Applied to experiments, robustness amounts to checking whether the behavioural regularity found with the set of inputs $x^n \cup x^{\infty - n}$ continues to hold under a somewhat different parametrisation of the environment. Robustness here applies not only to the inputs, x^n, but potentially also to those inputs that were disregarded on purpose in the experiment. Section 4.3, in Chapter 4, contains a number of examples of such robustness checks regarding omitted inputs in the dictator game example – the property rights over the endowment and the player's position in the game, the social distance between the two players, etc. Illustration 8.3 provides an example of robustness checks of laboratory evidence from the gift-exchange game regarding the inputs at stake in the target real-world situation. These additional investigations inform us about the degree of universality of the phenomenon: whether it continues to hold as an increasing number of relevant dimensions are included in the experiment.

Definition 3: inference.
As the neighbourhood of the inputs considered becomes wider, in both range and scope, the question of generalisability to the real-world situation of interest becomes closer to one of inference. The question is now whether the causal mechanism observed in a given context continues to hold in significantly different cases. The changes in inputs considered here are, e.g., the duration of the relationship or the type of subject pool. This

Illustration 8.3
The external validity of gift exchange at work

As described in Chapter 4, Section 4.5.2, the relevance of gift exchange to describe work relationships has been much disputed. To test the generalisability of the experimental results to actual work relationships, Bellemare and Shearer (2009) conducted a field experiment on worker responses to an exogenous gift. The main difference with previous field evidence was to implement the experiment in a pre-existing work situation, and to use a one-shot gratification – in order to closely replicate the treatment variable studied in the seminal laboratory experiments. The study took place in a tree-planting firm in Canada, in which workers are paid a piece rate according to the number of trees planted per day. The sample was a team of 18 planters and the observation period was seven working days, from a Thursday to the following Friday. On the second day of the experiment, workers received a surprise bonus of $80. The workers were unaware that they were taking part in an experiment. They were told the bonus was a gift from the firm, which decided to distribute some extra money due to very exceptional circumstances. Workers were provided with a detailed credible explanation of these circumstances and the gift was understood to be non-replicable. The aim of the experiment was to measure the impact of this gift on productivity – interpreted as evidence of reciprocity at work. To this end, an external data set was used providing information on the productivity of the same workers under non-experimental conditions. The data allows for the control of a number of environmental factors, like the day of planting or weather conditions. This comparison shows that average productivity significantly rose on the day of the gift by 118 trees per worker (average productivity during the experiment was 1075 trees per worker per day). This effect is robust to weather shocks and day-of-the-week effects. Interestingly, the bulk of the positive effect occurs only on the day of the gift, with no significant productivity rise on subsequent days. In addition, the number of completed years of work within the firm clearly plays an important role, suggesting repeated-interaction effects. These results confirm that workers positively respond to a gift, even in the realistic environment of a firm. The experiment also raises some questions about the profitability of gift-giving, in that the average additional revenue generated by the incentive was much smaller than the cost of the gift.

inference is a stronger form of external validity, as the differences here are more fundamental than those considered in robustness tests. Using the formalism developed in Chapter 4, the question is to what set of values of real-world inputs X^n can we generalise from what is observed with the experimental inputs x^n. This view of external validity is the most common one; it is also that which most often leads to confusing conclusions. Conclusions regarding the external validity of experimental results based on this definition require us to define the scope of generalisability that is required to match the real-world situation of interest. This choice is obviously subjective and results in cases for, or against, the external validity of experiments based on different views about the kind of inference that should hold. For example, while a 30-minute task can represent some (short-run) work situations, six hours is more suited for others. But is this enough? Would we instead want to see what happens over a full month? Or over two years or even longer? Depending on the answer, the external validity of the results from

the gift-exchange game will be judged differently. As we will see in the next section, the answer will also lead us to rely on different empirical strategies to assess the external validity of a set of experimental results.

8.1.2 External Validity and the Definition of the Target Parameters

The definitions above refer to increasingly strong versions of external validity as generalisability is extended to inputs that are increasingly different from those in the experiment. The further we move in this direction, the closer the question becomes to that of the multiplicity of the target parameter – which is arguably a matter not of external validity, but rather of the relevance of what is measured by the research question.

In the context of treatment effects (Section 3.2.3 in Chapter 3), for instance, two parameters have received much attention: the average treatment effect, which is supposed to show what would happen were the treatment to be applied to a randomly drawn individual from the reference population; and the average treatment effect on the treated, which applies only to those individuals who actually receive the treatment. As described in Chapter 3, the difference between the two lies in the possible heterogeneity of the causal mechanism. If individuals from the treated and untreated populations differ, not only per se but also in how they react to the treatment, then the two parameters will differ.

Obviously, we cannot in general measure the population average treatment effect from the response of a selected subpopulation. For example, a vast majority of participants in laboratory experiments are typically university students (this is a very common criticism of laboratory experiments, which will be discussed in Section 8.3.3). If students behave differently from the rest of the population for the tasks under consideration, the laboratory evidence might not be generalisable. As such, we identify an average treatment on the treated, as for any empirical analysis of a particular sample, when the sample and the rest of the population behave differently. It is important in terms of external validity to establish whether the sub-sample is representative 'enough' of the target population in the real world, i.e. whether the laboratory treatment effect is meaningful in real life. External validity is, then, not so much a question of the empirical method, but rather one of inference: to what extent is the causal parameter heterogeneous, and how large is this heterogeneity? This question is in no way particular to laboratory experiments, or even to experiments in general. It rather refers to the universal question of the representativeness of the samples in empirical work, the definition of the target parameters, and the choice of what should be measured to answer the research question. These questions are no different from those faced in field experiments (Do the results generalise to other time periods? To alternative definitions? To different magnitudes and implementations of the treatment? And so on) or the statistical analysis of natural data (see e.g. Heckman, 1996, 1997; Angrist and Imbens, 1999; Deaton, 2010).

The formal framework introduced by Falk and Heckman (2009) helps illustrate how narrow the boundary is between external validity and the precise definition of the target parameter. Denote Y the outcome variable of interest and $X_1, X_2, \ldots, X_J$ a list of all

determinants of Y, among which the effect of X_1 on Y is the main focus of the empirical analysis. Writing $Y = f_y(X_1, \ldots, X_J)$ as the true relationship determining Y, the causal effect is defined as $f_y(\Delta X_1, \tilde{\mathbf{X}})$: the change in Y induced by a change in X_1 holding all other factors $\tilde{\mathbf{X}} \equiv (\tilde{X}_2, \ldots, \tilde{X}_J)$ constant. The inference of this measure to other situations, associated with different values $\{X_1', \tilde{\mathbf{X}}'\}$, crucially depends on the assumptions about the function f_y.

- If f_y is assumed to be additively separable, such that $Y = f_{x_1}(X_1) + f_{\tilde{x}}(\tilde{\mathbf{X}})$, the causal effect will be the same whatever the value of the controlled inputs $\tilde{\mathbf{X}}$. The question whether these inputs are set to their real-world values is thus irrelevant.
- If the function f_{x_1} is further assumed to be linear, then the effect is also independent of value of X_1 at which the causal effect is observed.

In the general case, in which none of these assumptions hold, any empirical result will be particular to both the values of $\tilde{\mathbf{X}}$ and the neighbourhood in which X_1 changes. No empirical method (in particular, neither field nor laboratory experiments) can thus deliver universally valid measures of the causal effect, unless the above two assumptions hold – in which case all empirical methods are equally capable of achieving this goal.

To illustrate, consider the imaginary situation in Figure 8.1, first introduced by Leamer (1983). Figure 8.1.a shows the outcomes following from two different values, X' and X, of the determinant of interest, X_1. From this experiment, the causal change in Y resulting from $\Delta X_1 = X' - X$ can be inferred, but there is no way of disentangling the three underlying functions f_y displayed as the continuous, dotted and dashed lines (out of an infinite number of potential candidate relationships). As illustrated in Figure 8.1.b, this comes about due to a lack of data: if the relationship is observed a number of times, for a wider range of values of X_1, then the nature of the underlying relationship can be

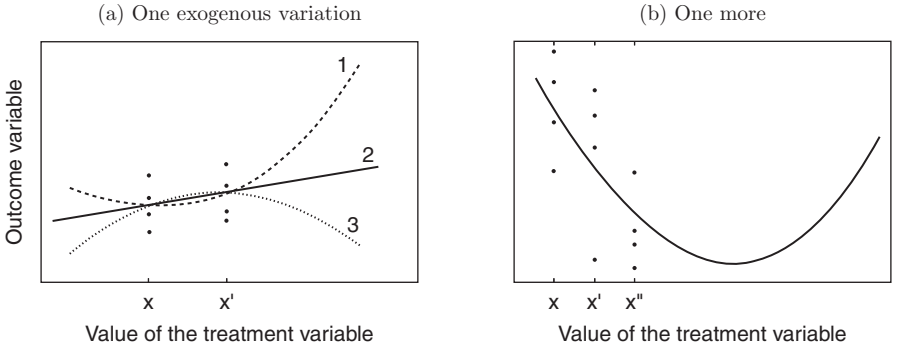


Figure 8.1 The identification of heterogeneous treatment effects

Note. Each dot in the figure shows the value of the outcome variable obtained in a (thought) experiment from the value of the target explanatory variable on the x-axis. The continuous lines illustrate the variety of true causal relationships that are consistent with the observed data.
Source: Leamer (1983, p. 35, Figure 1-2)

better determined. It also illustrates that the results from different methods each considering different neighbourhoods of X_1 (e.g. a lab experiment looking at $\Delta Y|_{\Delta X_1 = X' - X}$ and a field experiment looking at different values, $\Delta Y|_{\Delta X_1 = X'' - X'}$) will reach different conclusions, as the value of the treatment effect is different. This results from their different target parameters, all generated by the same, non-linear, true relationship.

When comparing the results from different methods, each measuring the causal effect of interest at different points in the distribution of controls $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}', \tilde{\mathbf{X}}''$, and/or in different neighbourhoods of the variable of interest, $X_1 = X, X', X''$, the choice mainly reflects the research question – which of these causal effects best defines the relevant target parameter? In terms of inference, it is also worth stressing that there is no clear advantage in choosing the measure delivered by $\{X, \tilde{\mathbf{X}}\}$ over that from $\{X', \tilde{\mathbf{X}}'\}$ (or vice versa) to infer the effect under conditions $\{X'', \tilde{\mathbf{X}}''\}$: this limit is a direct consequence of the data-generating process, and in no way depends on the method used to elicit the causal effects under $\{X, \tilde{\mathbf{X}}\}$ or $\{X', \tilde{\mathbf{X}}'\}$. In these circumstances, the only way to generalise the observed findings to alternative conditions is to obtain some knowledge of the function f_y – i.e. to combine empirical analysis with a theoretical model.

This point is illustrated in Figure 8.2, which considers a wide range of observed treatment effects at many different values of X_1. The true relationship here is very non-linear and non-monotonic. As a result, different values of the true treatment effect will

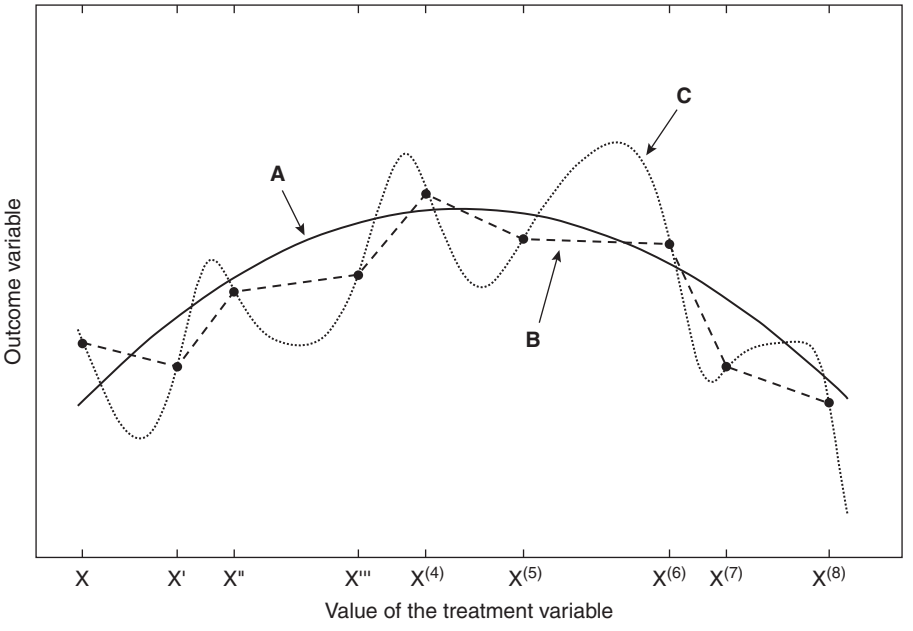


Figure 8.2 Many very heterogeneous treatment effects

Note. As in Figure 8.1, the dots shows the value of the outcome from a variety of values of the target explanatory variable, shown on the x-axis. The continuous lines illustrate the variety of true causal relationships that are consistent with the observed data.
Source: Leamer (1983, p. 36, Figure 3)

potentially be identified in the variety of situations observed. However complete these empirical observations, there will always be many ways of constructing a functional form describing the relationship which matches all the observed points. If we then want to pin down a particular candidate, some prior information is required.

8.1.3 The Wide Variety of Challenges to External Validity

Although the wide variety of definitions of external validity is with no doubt an impediment to the clarity of methodological discussion, there is little hope of converging on a unique definition that would universally apply to all laboratory experiments. This is the case as external validity has to be assessed differently depending on the scientific answer that is sought from the experimental evidence. From this point of view, the Roth (1988) classification discussed in Section 4.1, Chapter 4, provides a useful starting point.[3]

When testing theory, the aim of the experiment is to replicate the theoretical environment to assess whether the inputs do actually produce the theoretical prediction. It has long been recognised that the external validity of the experiment in this context is the same as that of the theoretical model to be tested (Plott, 1982). Theoretical models reduce and simplify reality in exactly the same way as do experiments, as this is the only way of being able to focus on cleanly defined mechanisms. This produces the same departure from the complexity of reality as in experimental design, and gives rise to the same doubts about the relevance of these mechanisms once the full complexity of the real world is accounted for.

There are two main consequences for external validity in this context. First, the burden of proof is reversed: the experiment can be seen as a complement to the theoretical insights, so that external-validity concerns apply rather to the model than to the experiment. Second, the question mainly applies to experiments that produce results supporting the theory, as only then do we wonder about the close relation of the theory to reality. As stressed by Plott (1991, p. 905) 'Models that do not apply to the simple special cases [implemented in the laboratory] are not general and thus cannot be viewed as such.' In other words, when models are supported by the simple cases analysed in the laboratory, the question of their generalisability is a natural further step. The laboratory environment works as a 'mediator', an intermediary stage in the incremental process from economic theory to its concrete application (Guala and Mittone, 2005). Schram (2005, p. 232) illustrates the point with the development of a new aeroplane: it is only if the plane does not crash in test-bed experiments that its use for passenger transport will be actually considered. But before getting to this stage, many further empirical investigations will be required to check the external validity of the results regarding naturally occurring flying conditions. On a final note, this line of reasoning also explains why external validity has long been seen as a second-order issue in the community: most early experiments aimed to test clearly defined theoretical models or mechanisms,

[3] As a reminder, Roth (1988) proposes a threefold classification of the objectives of laboratory experiments: testing theory, i.e. assessing the empirical relevance of theoretical models; searching for facts, in which case experiments use reality to inform theory; and whispering in the ears of princes, in which case experiments serve to improve the decision-making process by informing regulators or decision-makers.

putting aside the question whether the experiment itself provides relevant information about the real world.

External validity has become much more central, by contrast, as more and more experiments attempt to search for facts, which are strongly oriented towards real-world phenomena. When the aim of the experiment is to detect empirical regularities, external validity becomes a first-order question. Even if this makes external validity an important concern when dealing with experiments looking for facts, two nuances should be stressed. First, experiments that look for facts might be designed either to challenge or to inform theory – in which case external validity needs again to be considered in relation to the target theoretical framework. Many of the decision heuristics documented in Tversky and Kahneman (1974), for instance, are mainly aimed to reveal the limitations of rationality assumptions, with an eye on the consequences for the behavioural assumptions in theoretical models, rather than specific applications. Second, empirical regularities can also be insightful in their own right – independently of their external validity. Chetty (2015), for instance, notes that documenting the theoretical possibility of Giffen goods, even though goods with these properties are unlikely to exist in most actual markets, delivers useful insights into consumer-behaviour analysis. Similarly, documenting behavioural regularities, even in highly artificial contexts, provides useful guidance into the most salient behavioural assumptions, independently of how close they are to any real-world situations.

This view is also advocated by Camerer (2015), who labels it the 'scientific view' of laboratory experiments. From this point of view, a fact is a fact, and documenting individual responses to a well-defined environment helps us to better understand behaviour, whether or not a real-world counterpart to this environment exists ('The scientific view is that all empirical studies contribute evidence about the general way in which agents' characteristics, incentives, rules, information, endowments, and payoff structure influence economic behaviour,' Camerer, 2015, p. 251). Here, external validity, while not entirely irrelevant, is not the first-order concern when considering an experiment's contribution. The 'policy view', on the contrary, aims to use behavioural regularities observed in the lab to design real-world applications: to predict the effects of a new policy, to suggest alternative implementations of a given rule, to isolate the causes of unwarranted side effects, etc. It is only when the aim of the experiment is of this kind that external validity becomes key. Both this classification and that of Roth (1988) underline that the weight put on external validity depends on the nature of the research question – about what does the experiment aim to provide information?

In the same vein, an additional dimension to consider is whether quantitative measures or qualitative changes are the main aim of the empirical analysis. As stressed by Plott (1989, p. 1167), experiments predicated on the hypothesis that they were measuring numerical constants of nature 'would seem to require elaborate sampling procedures and explicit definitions of the populations to which the measurement is to be applied'. In many cases, however, experiments rather 'involve hypotheses about relative behaviour as opposed to numerical constants.' The external validity of qualitative results is generally easier to attain, as the contrary would require the effect of the move from the lab to the real world to be so strong as to reverse the sign of the observed relationship (Kessler

and Vesterlund, 2015). Going back to a distinction made in Chapter 5, Section 5.1.2, this often eases the generalisability requirement for experiments aimed at measuring treatment effects (i.e. changes in behaviour caused by changes in the environment) as opposed to measurement experiments.

8.2 Is External Validity Testable?

The previous section set out the definitions of external validity and how this depends on the research question the experiment addresses. However, one crucial question has been put to one side up to this point: how can external validity be addressed? Whatever the definition retained, external validity always amounts to the comparison of two empirical situations: one from the real world, consisting of an infinite number of dimensions that cannot be reproduced in the laboratory, and one in the laboratory, with very abstract environment and institutions. The question of external validity is not whether these two situations are different or not. They most surely are, as the very purpose of the laboratory is to overcome the lack of identification in naturalistic data by changing the environment. What is important is rather the relevance of the behaviour in the laboratory in the target situation, i.e. whether or not these differences produce fundamental changes in behaviour.

Once the essentially empirical nature of external validity is recognised, the next step is to say which empirical methods can be used to ascertain it. Siakantaris (2000) notes that testing the external validity of experiments yields a dilemma. There are two broad alternatives for testing external validity: the first uses experimental data and the second naturalistic data. Naturally occurring observations are in a sense the ideal benchmark for (and the only conclusive way to test) external validity, as what is observed comes from the real world. But, obviously, the point of alternative empirical methods such as experiments is the many limitations of such empirical investigations. When the research question is such that an experiment is required, it is likely that only experiments can ascertain the effect of any additional changes in the inputs. But if there are good reasons to doubt the external validity of experiments per se, we have a vicious circle: if an experiment aims to empirically establish the external validity of a previous experiment, the same doubts will also apply to this second experiment. This will equally apply to any subsequent additional experiments until the ultimate experiment replicates all of the dimensions of the natural environment in which the phenomenon under consideration occurs. But we already know that this cannot come about, as the real world cannot be transposed to the laboratory.

This is not only a seductive rhetorical argument. It highlights that external validity cannot in fact ever be *proven*. This echoes one of the conclusions of the discussion of internal validity in Chapter 5. In the same way as proper identification relies on the belief that there remains no confounding factor that is correlated with the measurement of interest, no definite proof of external validity will ever arise. There will thus always remain some room for methodological faith in favour of or against the external validity of laboratory experiments. The mechanical differences between the laboratory and the real world will always provide space for scepticism.

Illustration 8.4

Laboratory evidence of the external validity of declarative surveys

Attitudinal surveys are a useful tool to measure social attitudes in representative population samples. The World Values Survey (WVS), the General Social Survey and the German Socio-economic Panel are among the most commonly used sources in the literature. The World Values Survey has been run six times since 1981 across 80 countries. Social attitudes are measured through questions like 'Do you think most people can be trusted or that one needs to be very careful when dealing with people?', with Yes/No answers as a measure of trust. It is not clear whether these answers only measure intentions or actual actions: we might expect individuals to overstate their trust towards others, as compared to their real-life behaviour. Laboratory experiments, on the contrary, allow us to measure actions related to social attitudes (altruism, trust and reciprocity) via behaviour in games such as the trust game that was described in Illustration 4.5. A number of papers have combined these attitudinal questions with laboratory experiments to test the external validity of the former. Glaeser et al. (2000) first administered a survey on trust attitudes and behaviour to their student subject pool. Two weeks later, a subgroup of these students then participated in two different trust games. No correlation was found between declared trust attitudes and actual behaviour, so that general attitudinal questions are poor predictors of real choices. This seminal contribution was complemented by Johnson and Mislin (2012), who ran a trust game experiment in 35 countries with over 23,000 participants, divided into givers and receivers. The authors again find no correlation between the answers to the WVS questions and the amount sent by givers, which is a behavioural measure of their trust towards receivers. Interestingly, subjects generally appear to be more willing to trust than their survey answers suggested.

This does not, of course, mean that the investigation of external validity is a dead end. In the exact same spirit as the investigation of confounding factors, a natural empirical strategy is incremental (as described, e.g. by Guala, 2002). This process starts with the well-identified dimensions of the experimental environment that (i) distinguish it from its real-world counterpart, and (ii) are likely to influence behaviour to an extent that prevents the empirical research question being answered by relying only on experimental outcomes. The incremental process isolates the separate experimental dimensions and establishes whether the experimental results are robust to changing them: the following section will provide a number of different examples of such experiments. The validation process ends when the robustness checks have successfully been applied to all of the dimensions under suspicion. This is not a 'proof' of external validity, as the empirical evidence remains conditional on the inferential properties of the empirical method used to establish it. This should, however, suffice to establish the credibility of laboratory observations in reflecting the real-world phenomenon, as this is the most that can be done empirically.

If we take the incremental approach, the whole range of available empirical methods can be used to investigate the effect of incriminated dimensions under consideration. The preferred method will be the one that is best suited to this specific dimension – and laboratory experiments may be well suited for testing the external validity of non-laboratory measures, as shown in Illustration 8.4. A profusion of alternative empirical methods has emerged in recent years, in part instigated by the increased focus on external validity.

The list below provides an overview of these methods, ranked according to their distance from the real world, and also in descending order of the degree of control provided over the environment.

- **Laboratory experiments.** Although this may seem provocative at first glance, laboratory experiments can help with the incremental assessment of external validity by allowing additional relevant dimensions to be included in the experimental design. In the incremental view described above, criticisms of the external validity of laboratory experiments can be seen as a call for more experiments – with additional or different controlled inputs. The assessment of the external validity of results from laboratory experiments thus also makes use of two additional empirical methods: meta-analysis and replication, which are discussed in detail as a case study at the end of this chapter (Section 8.4).
- **Mobile labs.** These are laboratory experiments that are implemented outside the laboratory. A typical example is an experiment performed *in situ*, at the same place and with the same participants as the economic situation and agents under consideration (in a firm or school, for example). In practice, computerised experiments are run using tablets or laptops brought by the experimenter. The usual implementation rules (described in Chapter 6) have to be adapted to the particularity of the setting. For instance, preservation of anonymity via walls between participants, avoidance of communication between participants before and after the experiment is carried out, and control over interpersonal relationships between participants are all likely to be restricted relative to the laboratory setting.
- **Online experiments.** The ability to perform experiments online enriches the sample pool: first by adding more diverse participants in terms of individual demographics, and second because simpler implementation than in a physical lab allows larger sample sizes. This kind of experiment was first designed using online labour markets such as Amazon's Mechanical Turk (MT). This latter gives access to a very large pool of people who register for paid work, and hence allows relatively cheap large-scale experiments to be run. Existing work suggests that the respondents available on MT are more diverse than in typical experimental subject pools (Buhrmester et al., 2011; Paolacci and Chandler, 2014), although they are by definition a selected sample (in terms, e.g., of age, labour-force status or education). Paolacci et al. (2010) survey the internal-validity issues faced by social experiments on MT, Horton et al. (2011) discuss its use in economic experiments, and Crump et al. (2013) consider the particular case of cognitive measures. An alternative to the use of a commercial online labour market is the design of Web interfaces dedicated to online experimentation. In both cases, there is a loss in control due to the online setting, in which instructions cannot be explained face-to-face to participants, and decision-making takes place at home without control over either the time it occurs or what subjects are doing while they participate. Hergueux and Jacquemet (2015) provide a detailed discussion of how the design of experiments can be adapted to such a setting. The existing evidence based on comparisons between the lab and online implementations of the same experiment, of which Table 8.1 provides an overview, suggests no noticeable difference between the two.

Table 8.1 In-lab versus online experiments: overview of experimental comparisons

Type of experiment	Subject pool	Main results
	Anderhub et al. (2001a)	
Individual-level consumption/saving decisions	47 in lab, 50 online	(i) Similar economic behavior on average, (ii) greater behavioural variance online, (iii) shorter decision times online
	Shavit et al. (2001)	
Individual lottery-evaluation decisions	65 in classroom, 70 online	(i) Lower risk aversion online, (ii) greater behavioural variance online
	Charness et al. (2007)	
Lost-wallet game	178 in classroom, 124 online	Very little difference in average economic behaviour
	Fiedler and Haruvy (2009)	
TG with pre-play communication	136 in lab, 216 online	Lower levels of trust and trustworthiness online
	Chesney et al. (2009)	
DG, UBG, PGG, minimum-effort game, guessing game	Resp. 30, 64, 32, 31 and 31 online	Behavioural results qualitatively in line with previous laboratory-based experiments
	Horton et al. (2011)	
Watershed experiment, religiously primed and unprimed versions of the PD	Resp. 213, 189 and 113 online	Behavioural results qualitatively in line with previous laboratory-based experiments
	Amir et al. (2012)	
Public-good game, dictator game, ultimatum game, trust game	189 per game online	Behavioural results qualitatively in line with previous laboratory-based experiments

Note. For each paper in a row, the table describes the decision task, the subject pool and the main results from the comparison between an online implementation of the experiment and a standard laboratory implementation.
Source: Hergueux and Jacquemet (2015, p. 254, Table 1)

- **Field experiments**. Although this use of the empirical method is subject to many qualifications, as discussed in Section 3.5 of Chapter 3, field experiments allow the replication of experiments in environments as close as possible to the real-world phenomenon. As such, they are often considered to be the ultimate test of external validity. This test comes at the price of weaker control over individual behaviour – the accumulated evidence thus needs to be conclusive enough to be able to attribute any differences in the field to the setting itself, rather than to some uncontrolled dimension that is particular to the field experiment.

8.3 Testing External Validity

The incremental view of external-validity issues turns our attention to the features that render laboratory experiments different from the real world and lead to worries about

potential significant differences in behaviour. As an example, Levitt and List (2007, p. 154) identify five significant differences between lab experiments and real-world situations that may affect behaviour:

1. moral and ethical considerations,
2. the nature and extent of the scrutiny of one's actions by others,
3. the context in which the decision is embedded,
4. the self-selection of the individuals making the decisions and
5. the size of the stakes of the game.

A flourishing literature has empirically investigated these dimensions – see e.g. Camerer (2015) for a thorough review. One difficulty here is that the factors threatening external validity are likely partly application-specific (those that arise for the analysis of social preferences are likely different from those regarding risk or time preferences). Illustration 8.5, for example, describes an attempt to use external validity as a criterion to choose between measures of time preferences. It is part of the incremental process to define the features that are likely to be of first-order importance. It is important to underline that this means that no such collection of results can be thought of as proving the external validity of laboratory experiments per se. We do not here aim to answer this unanswerable question, but rather to highlight the doubts that are most commonly expressed about the external validity of experiments, and illustrate how accumulated evidence – making use of the whole scope of the toolbox described in the previous section – helps to narrow down the question.

8.3.1 Parallelism: Experimenter-Demand Effects

One of the first historically documented failures of external validity took place in management science with the Mayo (1949) experiment (later revisited by Levitt and List, 2011). This experiment took place in the Hawthorne plant of the Western Electric Company, near Chicago, in 1924. The aim was to measure the effect of better workplace lighting on performance at work. After observing a large performance improvement with the better conditions, many different changes were considered, each time producing the same effect on productivity. This was considered puzzling, until the reversion to the original lighting conditions was found to enhance productivity in exactly the same way. The conclusion from this experiment was that the fact of being observed was the main driver of the observed change in behaviour: workers complied with what they thought observers expected from them. This phenomenon is now known as the 'Hawthorne effect' in experimental social sciences.

This effect can be found in laboratory experiments if subjects do not change their behaviour according to the environment, but rather according to what they think the experimenter expects from them. This is an obvious challenge to the weakest possible definition of external validity, as outcome behaviour in the experiment will systematically differ from its non-experimental counterpart. A number of types of 'experimenter-demand effects' have been discussed in the literature. The first, scrutiny,

Illustration 8.5
The predictive power of experimental time-preference measures

To test the external validity of time preferences elicited in the laboratory, Chabris et al. (2008) correlated the real-world behaviour of their experimental subjects with their time preferences. They obtain mixed results, with time preferences being only weakly correlated with behaviour, but being at the same time the individual-specific covariate with the greatest explanatory power. Given the wide variety of methods available to elicit time preferences (as described in the case study in Chapter 6, Section 6.6), Burks et al. (2012) extend the research question to the preference-elicitation method with the best external validity. The first experiment aims to estimate the present-bias parameter and the discount factor in the quasi-hyperbolic (β, δ) model. The elicitation of preferences is based on the binary-choice method described in Section 6.6.1. The second measure elicits impulsiveness using a modified version of the design in Mischel et al. (1989), in which participants are asked to choose between waiting 10 minutes in front of a screen or pressing a button to reduce the waiting time and so renouncing a certain amount of money. The third measure uses a multiple-choice survey eliciting self-reported impatience among participants. The subject pool is made up of job trainees for whom a rich administrative data set of individual characteristics (e.g. demographic variables, cognitive test scores and financial variables such as family income and monthly expenditure) is available, along with six real-life outcomes: smoking, body mass index, the 'FICO credit score' (a measure of credit balance) and three job-outcome variables – whether participants leave their job during the 12 months following the training programme (LEAVE), whether they quit the training programme (WASH OUT), and whether they walk away from the job without resigning (i.e. is observed to be absent without leave, AWOL). The main results are summarised in the following table (from Burks et al., 2012, p. 318, Table 7), which displays the marginal effects of each measure on the behavioural outcomes in OLS regressions. All estimates include the individual characteristics as control variables.

	Smoking	Credit score	BMI	LEAVE	WASH OUT	AWOL
Present bias (β)	-0.099^{**}	0.014	0.025	-0.063^{*}	-0.080^{**}	0.045
Discount factor (δ)	-0.088^{**}	0.124^{***}	-0.001	0.031	-0.066^{*}	-0.102^{***}
Impatience	-0.014	0.038	-0.015	0.031	0.047^{**}	-0.006
Impulsivity	0.100^{***}	0.055	-0.009	-0.005	-0.019	-0.035
No. obs.	754	845	782	958	958	853

The standard quasi-hyperbolic (β, δ) model performs the best, as it predicts four of the six variables of interest. Present bias alone correctly predicts three outcomes, while the measures of impatience and impulsiveness predict only one outcome each.

is the same mechanism as that in the Hawthorne experiment. Participants react to the very fact of being observed and may try to match the experimenter's expectation that they will react to the design. The same issue can arise with cues about how to play the game that may produce social pressure on the subject from the experimenter. Consider,

for instance, experimental instructions that provide subjects with a full characterisation of the Nash equilibrium of the game that they are about to play, and a convincing case that this strategy is the only payoff-maximising rule of thumb. Most people will then seem to act consistently with the theoretical prediction, but the empirical evidence here is obviously poor and likely induced by the experimenter's manipulation of subject behaviour. Zizzo (2010) provides a full-length discussion of this phenomenon, with illustrations from the literature.

Second, a similar kind of issue arises if experimental subjects 'come to play' – the experimental set-up itself generates the desire to 'do' something, producing changes in behaviour that spuriously appear to be related to the environment (Carpenter et al., 2010). This might happen, for example, if subjects play a simple game many times over with no change from one period to the next: subjects may want to avoid repeating the same decision over and over again and make different choices just for the 'fun' of bringing about a change in their current experience. A last dimension to consider, in particular in experiments focusing on other-regarding behaviour, is the perceived wealth of the experimenter. Subjects are aware that any amount of money they earn counts as additional spending for the experimenter, and lower earnings saves research funds. This motive might not appear in the field – for instance, if applied to tax compliance, citizens may not care about government money to the same extent or in the same way as they do about the experimenter's research fund. Frank (1998) reports some reassuring evidence that subjects do not actually seem to care about the experimenter's wealth. Even so, the more weight is given to this aspect, the more likely it is that a lab-specific experimenter-demand effect will change behaviour. This would come about if, say, the experimental instructions make clear that any money saved via lower earnings would allow for further research on the topic or more credible scientific results.

All of these dimensions are clear examples of spurious changes in behaviour resulting from the relationship between the subjects and the experimenter. Beyond the obvious mechanisms described above, experimenter-demand effects are generally difficult to prove or fully distinguish from legitimate design choices. There are both an experimenter and subjects in a laboratory experiment, and the experiment itself is the result of their relationship: we can always suspect an experimenter-demand effect, as the experiment reflects experimenter demand in itself. Last, to a large extent, all economic situations involve some external authority (the state, the manager of a firm, the market authority, etc.) that is in charge of implementing the rules surrounding decision-making. The key question here regarding demand effects challenging external validity is to ask whether subject compliance with the experimenter's requests is different from compliance with the same requests from real-world authorities.

8.3.2 Robustness: The Effect of Laboratory-Specific Implementation Rules

As discussed in Chapter 5, laboratory experiments aim to construct a microeconomic system focusing only on the subset of factors under consideration. Concerns over the internal validity of the measures lead to many implementation rules that seem strange in comparison to the real world. By nature, all of these attempts to establish

the identification of the target parameter (via greater internal validity) pose potential problems for external validity, as behaviour might be significantly different once these parameters take on their real-world values. This section provides a review of the features that are most commonly identified as being laboratory-specific.

The Size of the Monetary Stakes

The dominance principle (described in Section 5.2.1) leads most experiments to offer lower rewards than those in many economic situations. As the pool of subjects is typically mainly composed of undergraduate students, experiments often propose twice the minimum wage for one hour of participation. If the experiment concerns, e.g., investment decisions by firms or employment contracts based on long-term relationships, inference implicitly assumes the monotonicity of behaviour in the size of incentives. Although potentially costly, robustness can be checked using additional treatments with higher (or lower) incentives that better match the situation under consideration.

The literature review in Camerer and Hogarth (1999), described in Illustration 5.5, finds mixed evidence of the robustness of experimental results to the size of monetary rewards, although Holt and Laury (2002) and Lefebvre et al. (2010) uncover large movements in risk attitudes as the size of the stakes increases. A recent research trend overcomes the budget constraint associated with these robustness checks by changing the opportunity cost of the subject pool rather than the value of the stakes: the experiment is run in countries with lower market wages – typically developing countries. For example, Andersen et al. (2011) analyse the ultimatum-bargaining game in poor villages in north-east India with monetary amounts corresponding to between 1.6 and 1,600 hours of work. Figure 8.3.a depicts the observed behaviour of senders and Figure 8.3.b that of receivers, both as a function of the monetary stakes (measured

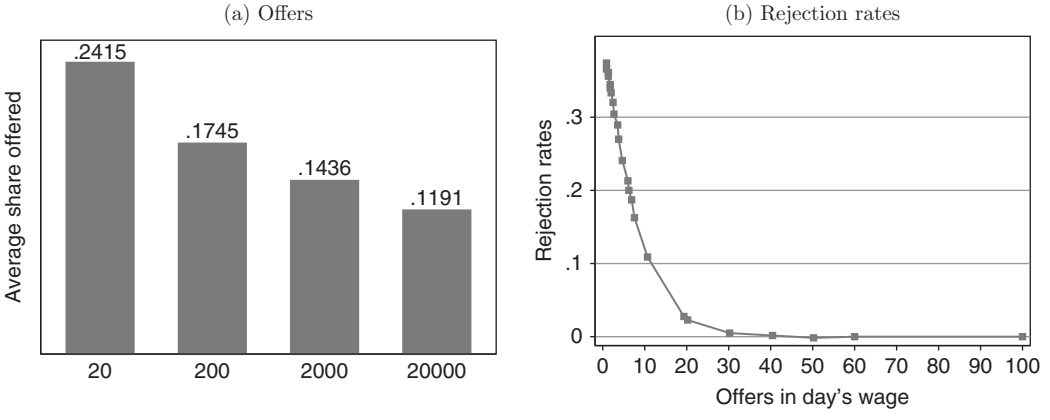


Figure 8.3 Social preferences when the monetary stakes are (very) high

Note. These figures show the share of the endowment offered by senders (*left-hand panel*) and the rejection rates from receivers (*right-hand panel*) as a function of the monetary amounts measured in Rupees.

Source: Andersen et al. (2011, p. 3431, Figure 1, p. 3434, Figure 3)

in local currency). Senders' behaviour does depend on the size of the stakes – higher stakes reduce the share sent. But the most important difference compared to the standard stylised facts regards receiver behaviour: the rejection rates quickly drop to 0 as stakes rise: the willingness to pay to punish unfair offers falls with the price of the punishment.

The source of the monetary endowment

When subjects are required to spend some money in the experiment (e.g. by giving money in a dictator game), it is common to endow subjects with windfall earnings at the beginning of the experiment. This obviously stands in sharp contrast to most economic situations, in which people spend money they previously earned as the reward for some activity. External validity will be threatened if the property rights over the endowment affect decisions. In this case, participants might act as if they are playing with 'easy' house money (Thaler and Johnson, 1990; Cardenas et al., 2014).[4] This question can again easily be tackled via additional laboratory experiments in which the endowment is earned rather than being a windfall. Money is earned by having subjects participate in a preliminary task for which performance is compensated. In the dictator game example, the evidence in Cherry et al. (2002), discussed at length in Section 4.3.2, shows a considerable difference between subjects splitting what they see as earnings from a general-knowledge quiz rather than windfall endowments. Oxoby and Spraggon (2008) generalise the approach to receivers' earned wealth, and again find a large effect: offers rise compared to the benchmark with windfall endowments on both sides. Equally, Augenblick et al. (2015) show that elicited preferences over time change with monetary rewards and real effort. Their results suggest that money over time is fungible, whereas real effort over time is less fungible and closer to a consumption good.

Induced values and the artificiality of the choice task.

Induced values were first introduced in market-design experiments (see e.g. Section 1.2.1 and Chapter 2 for examples) in order to increase experimental control over individual preferences towards the good that is exchanged on the market. Greater internal validity again comes at the cost of an artificial decision environment. Here 'consumption' for buyers amounts to being the owner of an artificial good about which only some monetary value is known; producers similarly supply a good that is defined only as a face-value marginal cost, etc. External validity can be assessed by applying the same market mechanism to actual home-grown goods, as discussed, for instance, in Illustration 2.1 for a Vickrey auction and Illustration 8.6 for a public-good game.

This induced-value approach later became popular for the study of behaviour in a wide range of decision contexts, and in particular effort at work. As shown by the experiment described in Illustration 4.4, effort is reduced in these experiments to a simple

[4] In choice under risk, Thaler and Johnson (1990) find that house money increases risk-taking in gains, whereas Cardenas et al. (2014) find less support for the house-money effect in an experiment involving both gains and losses.

Illustration 8.6
External validity of free riding in voluntary-contribution mechanisms

To assess the external validity of the voluntary-contribution mechanism game as a measure of free riding in public-good situations, Goeschl et al. (2017) explore whether the cooperative behaviour observed in the laboratory generalises to a climate-change public-good game. The experiment is based on two tasks: a *real-contribution task* and a *laboratory public-good game*. In the real-contribution task, participants are given the possibility of using a share of their show-up fee (€10) to reduce CO_2 emissions by buying (and destroying) emission permits from the European Emission Trading System (EU ETS). Participants are told the amount of CO_2 reduced per euro as well as the estimated delay in the resulting beneficial impact on climate change. In the second task, subjects are randomly matched into groups and play ten independent one-shot public-good games, each with a different combination of parameters (group size, marginal per capita return and payoff structure). Two distinct pools of players participated in the experiment: 43 students and 92 individuals from the general population. External validity is measured by the correlation between the real-contribution task and the public-good game. The main results are shown in the following Table (from Goeschl et al., 2017, p. 9, Table 3, and p. 16, Table 4), with the ten independent one-shot public-good games indexed *a* to *j*.

Decision	Group size	Symmetric	MPCR	**Correlations**		
				Non-students	Students	Whole sample
a	Large	Yes	0.10	0.0270	0.1689	0.0985
b	Large	No	0.10	0.1081	0.3723**	0.1822**
c	Large	No	0.15	0.1319	0.3516**	0.2003**
d	Large	Yes	0.20	−0.0184	0.2939*	0.0737
e	Small	No	0.33	0.0906	0.2964*	0.1713**
f	Small	Yes	0.40	0.0827	0.1455	0.1404
g	Large	No	0.42	−0.0074	0.0570	0.0446
h	Small	No	0.46	0.0242	0.1880	0.0956
i	Small	No	0.53	−0.0452	0.1308	0.0042
j	Small	Yes	0.80	−0.0719	0.1138	0.0491

The comparison between the subject pools shows that students' decisions are more consistent (correlated) than are the non-students' decisions. However, the correlations between average contributions are rather small in both tasks for both samples. The table also shows considerable changes in the correlations according to the experimental set-up. For example, the correlations turn out to be higher when the marginal per capita return is lower and the group size is larger.

cost-benefit pair – the cost being paid by the employee for the benefit of the employer. The external validity of the behaviour observed in this artificial task can be tested using 'real-effort tasks', in which performance depends on some individual skill – solving mazes, adding up numbers, etc. (Charness and Villeval, 2009). This setting is closer to the real-world idea of effort (see e.g. Gill and Prowse, 2015, for both an example and a survey of existing types of real-effort task). This comes with a loss of control, since the abilities that underlie individual performance are now private unobserved information.

For instance, Takahashi et al. (2016) emphasise the importance of whether the task is boring or interesting for subjects. Two different kinds of question arise regarding the generalisability of the experimental results based on effort tasks. The first is whether real effort changes behaviour relative to the induced-value setting. Bruggen and Strobel (2007) make this comparison in the context of an employer–employee gift-exchange game. They do not find any effect of switching from an induced-cost function to a real-effort task based on mental calculations performed in a limited amount of time. Lezzi et al. (2015) obtain similar results in a contest game. If real effort does indeed change outcomes, the second question then turns to the choice of the real-effort task that best matches the relevant real-life situation. For instance, Dutcher et al. (2015) distinguish tasks according to whether performance actually produces some wealth outside the experiment (*useful tasks*) or not (*trivial tasks*).

The neutrality of the decision context

The decision context is a highly sensitive part of the experimental design and is thus typically very specific to the laboratory. One of the rationales for this choice is to make the decision context clear and simple, which by itself could affect behaviour – see Illustration 8.7 for a robustness check about coordination behaviour. But the choice of the decision context is particularly important for the analysis of morally loaded behaviour (Cubitt et al., 2011; Masclet et al., 2003). Using words such as 'cheating' or 'lying' in the framing of the experiment is expected to change the extent of such behaviour, possibly affecting the generalisability of the quantitative measures obtained from neutral instructions in the laboratory. This question is addressed by Abbink and Hennig-Schmidt (2006) in the context of corruption. Neutral instructions reduce bribery to a game in which one player tries to affect another player's decision by a monetary transfer. In a separate treatment, this same environment is described to subjects using the words found in real-life bribery situations, leading to corruption rates that are quantitatively equivalent to those observed in a neutral context. If the aim of the experiment is rather to measure qualitative changes in behaviour, the context will matter in terms of external validity if it alters the way in which behaviour changes react to the environmental factors.

Tax-compliance experiments provide a number of examples of attempts to measure context effects. A tax-evasion game typically reduces compliance to the decision to report the amount of an endowment to later be taxed by the experimenter (see e.g. Torgler, 2002, for a survey). In this setting, the amount evaded is usually randomly fined. The game can easily be framed as a tax-simulation exercise to enhance the external validity of compliance behaviour. Choo et al. (2016) use this change in design to investigate the different compliance rates for student and non-student populations. They find no difference in a neutral context, in which the game reduces to lottery choices, suggesting that norms of compliance are an important driver of the difference observed in a contextualised tax-evasion game. Cadsby et al. (2006) similarly find that the response to the experiment's economic parameters strongly depends on the wording of the taxation exercise, with compliance elasticities becoming very small when reporting is framed as obeying an authority's request.

Illustration 8.7

Overcoming coordination failures thanks to complexity

Coordination failures are experimentally widespread in classical coordination games (see Chapter 5, Illustration 5.17). Experimental games, however, reduce coordination to its essential features – actions and payoffs – which arguably make the problem simpler and focus attention on coordination issues. To assess the robustness of coordination behaviour observed in the laboratory to more complex coordination situations, Parkhurst and Shogren (2005) compare behaviour in two strategically equivalent versions of a coordination game. The first version describes the actual target situation, in which farmers must choose simultaneously and independently which part of their land they give up. The payoffs of the resulting outcome are defined according to the objective of creating an adjacent protected area. Overall, the game features four players and more than 68,000 possible choices, resulting in $68{,}000^4$ possible outcomes, among which more than 9,000 are Pareto-rankable – and only one is Pareto-dominant. The second version is a simplified coordination game, in which the situation is reduced to two players and eight decisions. For each of the two versions, subjects play 20 rounds of the coordination game in quasi-stranger design. The table below (from Parkhurst and Shogren, 2005, p. 451, Table 1B) summarises the main results.

	Normal form		Grid game	
Rounds	Payoff-dominant	N	Payoff-dominant	N
1–5	82.2%	80	35.0%	80
6–10	90.0%	80	88.8%	80
11–15	90.0%	80	100%	80
16–20	93.8%	80	100%	80

Repetition clearly helps coordination on the payoff-dominant action in both situations. Complexity does impede coordination in the first few rounds, with more than twice as many payoff-dominant actions in the simplified coordination game. The situation is, however, reversed as time passes, coordination on the payoff-dominant action being achieved even more often in the complex game than in its simplified version.

The duration of the experiment

For practical reasons, for both subjects and the researcher, a typical experiment lasts between one and two hours. This is, of course, too short a period of time to be able to safely replicate long-term economic relationships. It is worth noting that there are some important exceptions to this rule. In particular, experiments that aim to elicit time preferences in a longitudinal setting need to be spread out over several sessions (Sayman and Öncüler, 2008; Halevy, 2015). This introduces complications in the design, as described in detail in Section 6.6. Another example comes from experiments in which the rules are too complicated to be understood by subjects in the usual time period. Here the experiment can be spread over several sessions, one in which instructions are described along with practice periods, and another eliciting actual decisions (see e.g., Stranlund et al., 2014, for an application to emission-permit markets).

These designs produce a loss of internal validity due to the time elapsed between the time slots, with, e.g., a risk that subjects interact regarding the experiment outside the laboratory, or look for information on how to behave. The implementation of the whole process within one single time period is therefore generally preferred. The field evidence on the gift-exchange game presented in Chapter 4, Section 4.5.2, shows that the duration of the experiment can drastically change the results. Normann et al. (2014) provide an extreme test of this effect in the context of a duopoly experiment. The main treatment of interest replicates a baseline two-player competition game, but spread out over one month, with subjects deciding only once on each day. Two versions of the game are considered, with firms endowed with either the same or different cost functions. The results show no difference in either behaviour or market outcomes when the cost functions are symmetric, and differences of only small size when they are asymmetric.

8.3.3 Inference: Subject-Pool Biases

Experiments are often run by faculty staff, who interact frequently with university students. Students, moreover, are intellectually and cognitively competent, are flexible in terms of their schedule, and generally have a low opportunity cost of time. For all of these practical reasons, a large fraction of laboratory experiment subject pools are made up of university students. One obvious challenge to external validity concerns situations in which (i) the real-world economic agents are not university students, and (ii) there are reasons to believe that student behaviour differs from that of the general population. Before discussing these two conditions, it is worth stressing that external validity is challenged only if they apply simultaneously to an experimental design. There is nothing to worry about if either students are the target population for which the results should be applied (observations on students are obviously solid regarding their own behaviour), or if there is no reason to expect their behaviour to be different from that of anyone else in the context of the research.

Students differ from the general population in a number of ways (see e.g. Slonim et al., 2013), just as any particular sub-sample of economic agents used in empirical work do (people working in a particular firm, consumers of a particular product, people living in a particular geographic area). The question in terms of external validity is thus whether these characteristics are likely to systematically influence behaviour, so that inference to other populations, with different values of these characteristics, can be called into question. In this respect, the distinction between measurement experiments and experiments aimed at measuring treatment effects (as described in Section 5.1.2) is again useful: representativeness is more likely to be an issue if the experiment aims to provide quantitative assessments of individual characteristics such as preferences or beliefs. Representativeness might, however, extend beyond measurement experiments if treatment effects are heterogeneous. Among student-specific characteristics (see Ball and Cech, 1996, for a detailed discussion), the most often discussed are the narrow age range and the low level of income as compared to the general population (e.g. Choi et al., 2014). Representativeness is a serious issue, but only if the differences in behaviour go beyond these observable individual characteristics: otherwise it suffices to control for composition differences to infer the likely average behaviour in any other

sub-population. Bellemare and Kröger (2007), for instance, find significant differences in behaviour in a gift-exchange game when comparing a representative sample of the Dutch population to the behaviour of experimental subjects. The lower levels of both investment and returns observed in the latter group, however, vanish once individual characteristics, such as age and gender, are taken into account.

Participants in experimental economics are also often enrolled in economics programmes. The choice of field of study may well be correlated with characteristics that influence behaviour in the context of a particular experiment. For instance, Frank and Schulze (2000) find evidence in a bribery experiment that economics students are more corrupt than students in other disciplines, and other people in general. An additional issue is the possibility of self-selection in the experiment, even within the selected sample of students – see Illustration 8.8 for an example. Anderson et al. (2013) perform a two-step comparison of other-regarding preferences: between self-selected students and self-selected adults, and between the latter and non-self-selected adults. They find differences according to the sub-population (students versus adults), but no difference resulting from selection.

Another particularity of the traditional experimental-economics literature is that the vast majority of experiments have been run in Western countries, where there is more investment in higher education, and easier access to facilities and research funds. In an influential paper, Henrich et al. (2010) wonder if the behaviour of 'WEIRD' people, from Western educated, industrialised, rich, and democratic countries, can be generalised to other kinds of population. Their work covers a wide variety of standard experiments (from many fields, well beyond economics) carried out in different parts of the world. For example, Figure 8.4 reports the results from two classic games

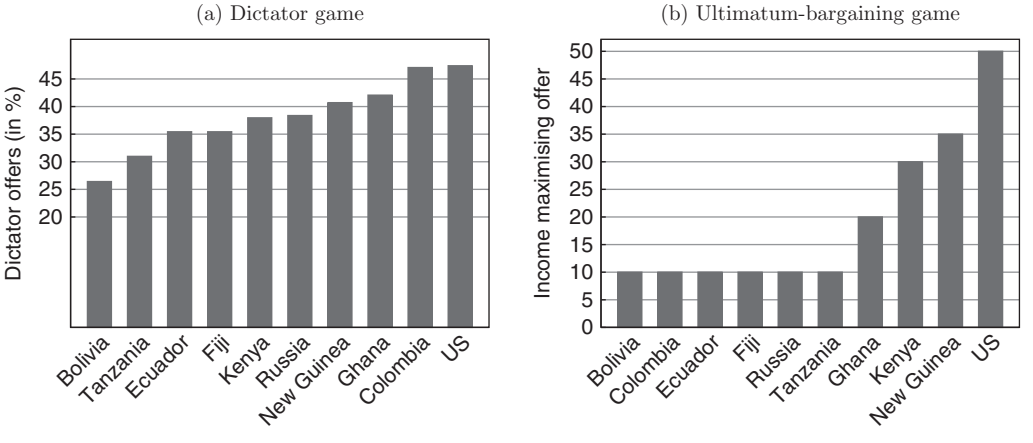


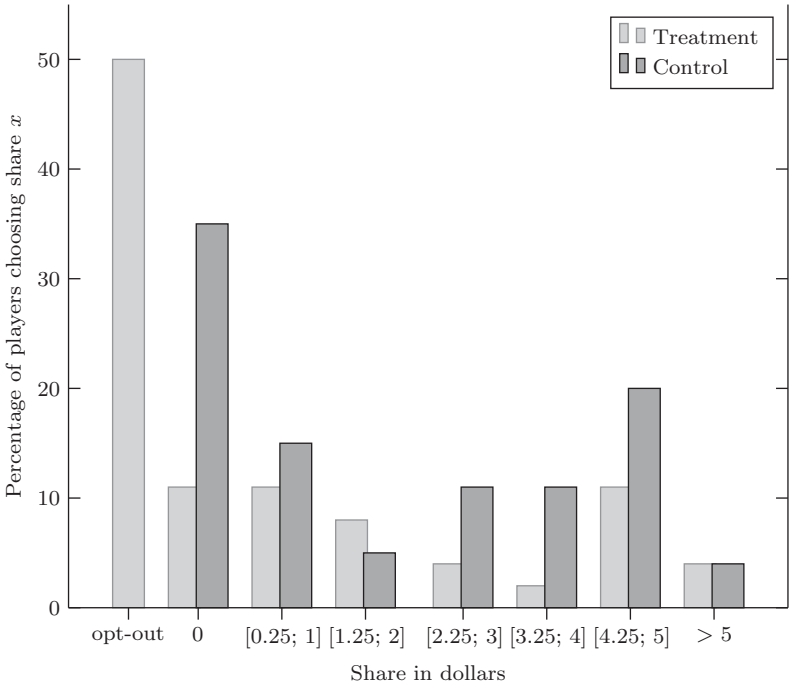
Figure 8.4 Other-regarding behaviour in non-WEIRD populations

Note. These figures show the share of the endowment offered by proposers in the dictator game (left-hand panel) and income-maximising offers in the ultimatum-bargaining game (right-hand panel) in experiments run in different countries.

Source: Henrich et al. (2010, Figure 3, p. 6).

Illustration 8.8
Self-selection in laboratory experiments

Participants in laboratory experiments are usually unaware *ex ante* of the game that they will play. This then forces individuals to make decisions in economic environments that they might avoid in real life. Lazear et al. (2011) provide evidence of the consequences of this phenomenon in the context of a dictator game in a between-subject experiment. Participants in the control group play the usual dictator game with an endowment of $y = \$10$ to be shared. Participants in the treatment group play a modified dictator game, where subjects can opt out of the game. If they opt out, they receive an amount $y_{out} < y$. The figure below (from Lazear et al., 2011, p. 144, Figure 1A) shows the results from the experiment run at University of California, Berkeley (additional experiments were run in other locations as a robustness check).



At least half of the players in the treatment group chose to opt out and receive y_{out}. The aggregate amount shared by dictators was consequently significantly lower than in the control group. In particular, the share of people who gave nothing fell from $1/3$ in the control group to $1/10$ in the treatment group. The figure above shows that over 25% of respondents would share a positive amount in the control group but opt out in the sorting group. These 'reluctant sharers' share – and even generously so – if they have to, but prefer to avoid the sharing task. A second experiment tests the price sensitivity of opting out via a number of dictator games in a within-subject design. The first condition is NO OPTING OUT and the second proposes OPTING OUT with $y_{out} = y$. The last part of the experiment consists of a series of increasing

opt-out amounts of y_{out} – from 5% to 100% of y. The aim is to induce reluctant sharers to re-enter the game so that the total shared amount increases, while the conditional average shared amount falls. When sorting is costless, 46% of people enter the game, sharing on average \$2.88. A one-dollar subsidy for entering increases entrance to 76%, but the average amount shared falls to \$2.22. Interestingly, subjects who first re-entered the game are those who share less in NO OPTING OUT. This is not in line with the idea that generosity in the standard dictator game is a good predictor of the utility of sharing.

on other-regarding behaviour: the dictator game (in Figure 8.4.a, which delivers a measure of generosity to others) and receiver behaviour in the ultimatum-bargaining game (summarised as the income-maximising offer, i.e. the offer that maximises expected earnings given the observed distribution of refusals from the receivers, providing a measure of social norms). The behaviour elicited in the US is shown in both cases as a typical WEIRD benchmark. In both respects, the other-regarding behaviour in the US is a clear outlier. The main conclusion from the paper is that behavioural science needs to widen the scope of its experimental pools to ensure the satisfactory generalisability of the observed regularities. These observations led to great interest in cross-cultural comparisons in the literature, to assess the robustness of standard results to alternative populations. This is particularly welcome when the observed regularities reflect cultural norms or habits in which significant differences can be observed across the world. The main challenge in running this kind of cross-cultural study is to ensure the comparability of the experiment across locations. The monetary incentives are typically converted using purchasing-power exchange rates. The experimental instructions are generally translated to the other languages by a professional, and then translated back to the original instructions. Any difference is resolved, and the translation process is carried out again until convergence is attained (Brislin, 1970).

The external validity of experimental results is not only affected by the particularities of the subject pool, but also by any specific characteristics of the pool of economic agents to which the results apply. If student behaviour is representative of the general population, but the people in the target situation differ significantly from the general population, student behaviour will not inform us about the target parameter. The typical example here is experiments that aim to document behaviour for a particular professional occupation or economic sector. Montmarquette et al. (2004), for instance, compare decisions about a company merger taken in a real-task experiment by managers in a large company with the decisions in the same situation made by students from a standard subject pool. They find very similar outcomes, although the strategies differ significantly. Fréchette (2015, 2016) provides a systematic review of the literature comparing student behaviour to that of both professionals from the relevant industries and other population groups. There is in general little difference between the two, although there are some exceptions (e.g. Palacios-Huerta and Volij, 2008, in a zero-sum game).

When professionals' behaviour significantly differs from that of students, it is important to understand what may lie behind this discord. One hypothesis is that experience in the decision-making situation matters (this is an obvious, and significant, difference

Illustration 8.9

The winner's curse with experienced bidders

Dyer et al. (1989) provide an experimental test of the difference between experienced and inexperienced players in common-value offer auctions – in which players compete not to buy a good at the lowest price, but rather to sell at the highest price. Bidders receive an imperfect signal of the common unknown cost of producing the good, and send a sealed price to the buyer. The buyer buys the good from the bidder who announces the lowest price. The following table (from Dyer et al., 1989, p. 11, Table 1) shows the results. The first column shows the percentage of times that the lowest bid is submitted by the lowest-signal player, which is a measure of the consistency of strategies across players. The second column shows average profits. A negative profit corresponds to the winner's curse. The last column shows the average profit in the theoretical equilibrium.

Pool	% of auctions won by the low-signal bidder	Average profits	Average profit in the theoretical equilibrium
Naive	71	-0.37	5.02
Executive	79	-1.01	5.42

There are no significant differences across populations in any of these indicators. The auctions are won 71% of the time by the low-signal bidder with 'naive' inexperienced players and 79% of the time with the executives. Profits are negative or almost 0 in all experiments. In addition, the share of bids under the expected cost is 66% for the naive players and 67% for executives. From these results, Dyer et al. (1989, p. 113) conclude that 'the winner's curse phenomenon is robust across auction form, market size and subject population'. Of course, executives could not systematically make these losses in their industry, otherwise their firms would not survive. This robustness result shows that experience cannot explain the behaviour observed in the laboratory.

between professionals and the general population). This hypothesis goes well beyond any comparison between professionals and others, and applies to any economic situation in which agents are frequently exposed to the same choices. Experience has been shown to reduce market anomalies in selling and buying behaviour (List, 2003, 2004). Illustration 8.9 provides another example relating to the winner's curse. Experience has also been identified as the main source of variation when comparing sample pools in labour economics – see e.g. the survey in Falk and Fehr (2003), showing that experience is more influential than socio-economic identity.

8.4 *Case Study*: Replication: Enhanced Credibility Thanks to Accumulated Evidence

A last dimension that is discussed in the context of the validity of experimental results is the size of the sample in published work. Bellemare et al. (2014) report that among the 71 papers published in the 2012 and 2013 issues of the journal *Experimental Economics*,

the range of independent observations per treatment was between 13.5 and 420, with a mean of 77 and a median of 51. Increasing the size of the experiment sample, although costly, is as simple as adding more experimental sessions and inviting more experimental subjects. This can be done in the course of a single experiment, or accumulated from a number of experiments with the same experimental environment. The independent observations in the set of all such experiments is a meta-sample the size of which grows as more experiments are run. This raises the question whether the experimental environments actually are the same, and if they are not exactly the same whether they deliver similar-enough information to be seen as part of the same meta-sample. These questions are exactly those raised in the context of a growing concern in empirical economics in general, and experimental economics in particular: replication, which is seen as a way to enhance the credibility of empirical results.

8.4.1 'Keeping the Con out of Experimental Research': the Need for Replication

To simplify the exposition, we here frame the discussion in the particular context of experiments aiming to detect a treatment effect – i.e. to know whether a change in one dimension of the environment has a statistically significant effect on the outcome.[5]

The concern over sample size is a matter of statistical inference. As described in Chapter 7, Section 7.3.1, decisions made on statistical tests (the statistical significance of the difference in the outcome) are susceptible to two kinds of mistake: false positives (type I errors, denoted α), and false negatives (type II errors, denoted β). As the two cannot be minimised simultaneously, the common practice is to apply the Neyman principle and arbitrarily fix the level of the test, α, which gives the probability of a false positive – the probability that the statistical test rejects the null hypothesis even though the null actually describes the data-generating process. Since this probability is embedded in the definition of the test, it does not depend on the size of the sample, but rather applies to resampling. This describes the asymptotic distribution of false positives in the population of experimental results that are due to the very nature of the statistical test. The type I error is thus in itself a first reason why replication is required to establish the credibility of empirical results. Out of a large number of identical attempts to measure the same treatment effect that does not actually exist, $\alpha\%$ will manage to find an effect. As this argument is statistical, we never know whether the first appearance of a result should be interpreted as (good) luck, or rather as an actually significant underlying effect – only accumulated and consistent evidence will make it increasingly unlikely that significant treatment effects are false positives.

This discussion applies to the repeated use of the same test, in particular with the same value of β, the type II error, which measures the likelihood of a false negative – the probability of failing to find an effect that actually exists. The Neyman principle leads us to use the most powerful (if it exists) of the consistent tests, where a consistent test is such that the power, $1 - \beta$ approaches 1 as the size of the sample goes to infinity. In contrast to type I errors, this property of the test thus does depend on the sample

[5] The title of this section is borrowed from two early contributions in this context, Leamer (1983) and Roth (1994).

Table 8.2 Calculation of the false-positive report probability

Truth of the alternative hypothesis	Significance of test		
	Significant	Not significant	Total
True association	$(1-\beta)\bar{y}$ [True positive]	$\beta\bar{y}$ [False negative]	$\bar{y}$
No association	$\alpha(1-\bar{y})$ [False positive]	$(1-\alpha)(1-\bar{y})$ [True negative]	$1-\bar{y}$
Total	$(1-\beta)\bar{y}+\alpha(1-\bar{y})$	$\beta\bar{y}+(1-\alpha)(1-\bar{y})$	$\bar{y}$

Note. Given the true relationship in the rows, the cells show the probability of significant (first column) and insignificant (second column) results in the literature according to the level (α) and power of the statistical tests (β) as well as the share of true associations among those investigated, $\bar{y}$.
Source: Wacholder et al. (2004, p. 440, Table 1).

size. Moreover, the likelihood of a type II error is higher the smaller the difference between the null hypothesis and the actual data-generating process. In other words, the likelihood of a type II error is larger the bigger the sample size needed to achieve a given power (see Focus 7.9 for an illustration of how the sample size is adjusted according to power).

All empirical work is subject to both false positives and false negatives, but their implications for the credibility of the results published in scientific journals exceed their face value. Wacholder et al. (2004) develop a method to calculate the probability that a positive finding picked up in published research is actually false (see Maniadis et al., 2014, for a transposition to experimental economics). This refers to the probability that the null hypothesis is true given that a significant effect is found, $Pr[H_0|d_1] = Pr[\theta_0 \in \Theta_0|\mathbf{T}(\mathbf{Y}) \in R]$ – which is the reverse condition as compared to the size of the test, calculated conditional on the null hypothesis being true. Table 8.2 shows the joint probability distribution of the conclusion of an empirical analysis and the data-generating process, denoting by $\bar{y}$ the proportion of all relationships that can be investigated in a field that are actually true. The 'false-positive report probability' (FPRP) is the likelihood of a false positive among all the significant results that are reported:

$$FPRP = \frac{\alpha(1-\bar{y})}{\alpha(1-\bar{y})+(1-\beta)\bar{y}}$$

Assuming that all tests are of the highest-level power ($1-\beta=1$), this probability is the same as the standard size of significance tests ($\alpha = 0.05$) only if half the relationships under investigation are actually true ($\bar{y} = 0.5$): it is lower (higher) if this proportion is lower (higher). As an extreme example, Wacholder et al. (2004) calculate an FPRP of slightly over 98% when $\bar{y}$ is 0.001 – i.e. one effect out of a thousand that are investigated is actually true.

Ioannidis (2005) extends this framework in two ways. He first considers competition between researchers in the quest for identification of each relationship. If n teams are

looking for the same empirical result, all of which work with the same level of power, then the probability of being the first to find a true relationship is $(1-\beta^n)\bar{y}$ and the probability of the appearance of the first false positive becomes $[1-(1-\alpha)^n](1-\bar{y})$. The resulting probability of a false negative among the 'original' results in the scientific literature is

$$FPRP(n) = \frac{[1-(1-\alpha)^n](1-\bar{y})}{[1-(1-\alpha)^n](1-\bar{y}) + (1-\beta^n)\bar{y}}$$

This increases with n as long as $(1-\beta) > \alpha$.

On top of these pure consequences of the design of statistical inference, the credibility of empirical results is also conditional on the spread of scientific malpractice. There are many causes of these biases. A first is deliberate malpractice, such as the false reporting of information and all imaginable kinds of fraud, sometimes referred to as 'questionable research practices'.[6] These practices are designed to produce significant results where none appear, trivially leading to a rise in the share of false positives. A less extreme practice leading to the same kind of consequence is to widen the set of outcome variables when the hypothesised relationships are not found in the data, until significant relationships appear (a practice often labelled 'data mining'). This again leads to an inflation of significant results in the literature, and more strongly so if the failure of the first hypothesis is not reported.

But even truthful academic practice can yield biases. The two most important are 'publication bias' and 'reporting bias'. Publication bias refers to the selection of scientific results in academic journals from the willingness of editors and authors to increase the impact of published research. Publication bias implies that innovative results are more likely to appear in the academic literature than are failures, and the more so the more surprising and unexpected are the results. The process of academic publishing is therefore likely to distort the distribution of results in the literature towards more false positives. Reporting bias, on the other hand, refers to the selection of the data reported in a research article. As stressed by Roth (1994), this risk arises in particular when empirical findings come from a sequence of independent trials – e.g. a first treatment is designed, and leads to unsatisfactory results, but provides insights into the design of a subsequent treatment, etc. As such, independent trials are natural parts of the scientific process, made up of trials and errors. But the question of which results are finally reported arises; if only the most convincing results appear, while the intermediary treatments do not, there is again distortion in the distribution of significant findings that are conclusive, since the omitted treatments are more likely to have produced non-results. This leads to the extreme 'file-drawer problem', consisting in silence over failures to obtain significant results but the public release of significant findings.

To quantify the effect of these biases on the credibility of empirical results, denote by B the share of the overall set of positive results that arises due to this bias. Of the

[6] See List et al. (2001) for empirical evidence on the extent of this in economics, and Stapel (2014) for the detailed narrative of a recent example in experimental psychology, the impact of which was such that Kahneman (2012) wrote an open letter arguing in favour of better-designed replications.

$\bar{Y}$ true relationships, a share $(1 - \beta)$ will then be accurately found to be true and a share $B \times \beta$ will be reported to be true due to the bias; equally, of the $(1 - \bar{y})$ zero relationships, a share α will mistakenly turn out to be true due to type I error, and another share $B \times (1 - \alpha)$ will result from the bias. The share of false positives induced by the bias is

$$FPRP(B) = \frac{(1 - \bar{y})[\alpha + B(1 - \alpha)]}{(1 - \bar{y})[\alpha + B(1 - \alpha)] + (1 - \beta)\bar{y} + B\beta\bar{y}}$$

Ioannidis (2005) shows that this probability rises in B when $\alpha < 1 - \beta$, in which case higher bias leads to a greater share of false positives for the given levels of power and size typically found in the field.

All of the above factors imply that the credibility of any empirical field of research is challenged by a number of factors, which include, but are not limited to, sample size. The assessment of the likely extent of these factors led Ioannidis and Doucouliagos (2013) to conclude that the credibility of empirical economics is likely not strong. Replication is the most obvious way of addressing these issues in order to enhance the credibility of empirical results: 'as a profession, our best defence against erroneous conclusions resulting from unreported or incompletely reported search is to encourage experimenters to follow up on one another's work' (Roth, 1994, pp. 287–8).

8.4.2 From Replication to Cumulative Research

One comparative advantage of experiments is their reproducibility. Nothing is simpler than running the same experimental treatment again, i.e. exposing different subjects to the same rules and environment to see whether the same behaviour results as in the first attempt (at least if researchers comply with the good practice of making public the original instructions that were used to run the first experiment once the paper has been published).[7] It is an open question, however, whether replication should be strictly restricted to the experiments with exactly the same design, or whether changes in the design should be included.

Strictly defined, replication aims to increase the size of the original sample. This raises two issues. First, this narrow definition would apply not only to the rules, parameters and control variables, but also to any dimension of the original experiment. It is difficult, for instance, to decide whether the individuals invited to participate in the replication experiment belong to the same target population. The range of variation that falls into the set of strict replication thus needs to be clearly defined. Second, narrowly defined replications increase the power of tests in the new sample consisting of observations from the two studies, but does not deliver additional information about whether the combined result can be generalised. As noted by Roth (1994, p. 288),

[7] This does not mean that experiments are less likely to give rise to false positives; arguably, as experiments provide an easy way of generating data, they may even be more likely to produce malpractice and competition effects.

Table 8.3 Replication versus robustness: a classification

Sampling methods reported in original distribution of the parameter estimates	Sufficient conditions for discrepancy	Types	Methods in follow-up study versus methods reported in original			Examples
			Same spec.	Same spec.	Same sample	
		Replication				
Same	Random chance, error, or fraud	Verification	Yes	Yes	Yes	Fix faulty measurement, code, data set, address sampling error, low power
		Reproduction	Yes	Yes	No	
		Robustness				
Different	Sampling distribution changed	Reanalysis	No	Yes	Yes/No	Alter specification, recode variables
		Extension	Yes	No	No	Alter place or time, drop outliers

Note. This table describes the main characteristics of replication (top panel) and robustness (bottom panel) follow-up studies.
Source: Clemens (2016, p. 3, Table1).

> when a carefully conducted experiment is repeated, the likelihood that the data will be similar seems to be high. But precise replication gives little information about robustness. What ultimately gives us our best indication of the robustness of experimental results is replication with some variation of experimental parameters and conditions.

This kind of replication is what Ioannidis and Doucouliagos (2013) label 'conceptual replications', in which the research question and the design of the experiment are defined in reference to the original experiment, but feature variations that aim to test its robustness.

In an attempt to address these issues, Clemens (2016) provides guidelines on how follow-up experiments can be classified as (i) replication studies; (ii) robustness checks, and (iii) original contributions, as compared to the original experiment. Two dimensions are key here. The first focuses on the target parameter that we wish to measure, in particular whether the results are expected to be the same or different. The second focuses on the distribution of the sample parameters – whether the estimate calculated from the new data is drawn from the same distribution as the original estimates. The first criterion distinguishes replication in a broad sense from follow-up studies producing original contributions: it is only when the same empirical findings are expected in the two studies that the follow-up is a replication. This definition is required to restrict replication to work that aims to screen the false positives that appear in the process of statistical investigation. The second criterion can be used to distinguish strictly defined replication, in which the sample parameters are drawn from the same distribution in the two studies, from robustness checks that aim to measure the same target parameter but on a different sample. Table 8.3 summarises the resulting classification and provides examples of the

kinds of change that can be found for each type of follow-up. Each of these two groups can be further split into two kinds of implementation. On the one hand, replications are 'Verifications' if they use exactly the same data as in the original study to which the same analysis is applied, while they are 'Reproductions' if the analysis is performed on new data, i.e. on a different sample drawn from the same population. Robustness checks, on the other hand, take the form of 'Reanalysis' if a different statistical analysis is performed on the same data, and 'Extensions' if this is carried out using different data.

These definitions do not include follow-up studies that introduce large enough distortions in the design for the expected results to change. The above sections provided many such examples, including, e.g., changes in the subject pool or the monetary stakes. These changes no longer target the same population parameter: they rather aim to test the generalisability of experimental findings to alternative contexts, i.e. to assess how universal the relationships documented within the context of a particular empirical framework are.

Whether it arises in the course of replications, robustness checks or such follow-up investigations, this process produces accumulated empirical evidence regarding the same research question, with differences in the empirical strategy ranging from almost none beyond the sample size to more significant ones like the content of the experimental instructions or the monetary stakes. We typically appeal to meta-analysis to aggregate this information. This is a set of statistical tools for the collection and analysis of data referring to the same phenomenon. While most of the methodological discussions of meta-analysis appear in fields where there is pressure for policy recommendations with a considerable degree of confidence, such as in epidemiology leading to the 'MOOSE' standard (Stroup et al., 2000), or health care interventions with the 'PRISMA' recommendation (Liberati et al., 2009), efforts have also been made in psychology (American Psychology Association, 2008, defining the 'MARS' standard) and economics (see e.g. Stanley et al., 2013, for a detailed checklist aiming to establish a publication standard). These various recommendations yield the following guidelines on how to undertake a meta-analysis.

1. **The research strategy.** This concerns the definition of the target parameter, i.e. the precise theory or empirical phenomenon to be investigated, as well as its empirical measurement, along with the definition of a common metric that will allow the effects to be measured on a common scale.
2. **Literature search.** This covers the choice of databases and keywords to be used to search for existing literature. A set of inclusion criteria splits studies that fall into the range of the meta-analysis from those that do not.
3. **Data collection.** The retained studies have to be coded in a uniform way, which also implies deciding what information to record from each study. This step will likely require the interpretation of some of the information in the original papers (as this information needs to be comparable between analyses).

The typical observation in the data set is the result of a statistical treatment applied to the original sample. The secondary analysis requires particular statistical tools to take into account, for example differences in sample size and standard errors between observations. Hedges and Olkin (1985) is a seminal description of the statistical methods

that can be applied. Brockwell and Gordon (2001) provide a comparison of the most common techniques, while Rosenthal and DiMatteo (2001) focus on applications in psychology.

8.4.3 Examples and Current Practice

Despite the growing case in academia for more replication studies, replications are unanimously recognised as being underrepresented in economics compared to other disciplines (see e.g. Duvendack et al., 2015, for a recent discussion and a historical perspective). One obvious reason is that follow-up work, precisely because it covers empirical findings that are already known, is both less exciting for researchers and less attractive for editors (whose editorial responsibility is to publish results that will attract the attention of the scientific community, for example via citations, which is less likely to come about with consistent replications, and who might be reluctant to engage in the struggle associated with negative replications of well-received results). Zimmermann (2015), for example discusses the negative incentives authors, editors and referees face regarding replication studies, and the case for a new journal dedicated to this question.

There is, however, a growing willingness to overcome this lack of replication. For instance, the Economic Science Association launched a new academic journal in 2015, the editorial policy of which (as stated on the journal policy page, accessed on 17 January 2017) encourages the submission of 'article types that are important yet underrepresented in the experimental literature (i.e. replications, minor extensions, robustness checks, meta-analyses, and good experimental designs even if obtaining null results)'. Even so, it is fair to say that replications of the above type remain scarce, representing under 7% of the articles that have been published in this journal to date (possibly due to a lack of submissions). Perhaps more promising is the recent launch of a number of open-access databases referencing replications, of which the Replication Network (https://replicationnetwork.com) or Curate Science (http://curatescience.org) are active examples. This growing awareness also leads to more careful consideration of good practice in other disciplines. The American Economic Association, for instance, now offers a pre-registration system, mainly aimed at field experiments, whereby the main features of an experiment can be registered before it takes place: typically, the treatment to be used and the outcome variables that the treatment is supposed to affect. Pre-registration aims to reduce both the file-drawer problem and data mining (see Olken, 2015, for a detailed discussion regarding field experiments). Coffman and Niederle (2015) argue that the benefits of this practice might not outweigh the costs, in particular in fields in which replication is easy. These pre-registration procedures are, however, now available, and sometimes used, for laboratory experiments – e.g. https://osf.io or aspredicted.org.

Given this situation, the vast majority of existing replications or robustness checks in experimental economics are stand-alone replications of a given experiment. Typically, an original experiment is replicated to serve as the baseline for a new treatment. Numerous examples are available in the literature for the most popular experimental games,

such as the voluntary-contribution, prisoners' dilemma and social-preference games.[8] Until recently, however, there were no examples in experimental economics of large-scale replication experiments like those in the sciences or psychology – e.g. Klein et al. (2014); Open Science Collaboration (2015). A notable exception is Camerer et al. (2016), in which all 18 between-subject experimental contributions published in the 2011–14 issues of the *American Economic Review* and the *Quarterly Journal of Economics* are replicated in the laboratory. Overall, 61% of the significant effects in the original work were replicated, and the average size of the replicated effects was 66% of their original value.

Summary

The concern for statistical inference based on experimental measures has led to design choices where all of the features of the situation of interest producing noise and/or confounding variations have been eliminated. The experimental setting is consequently far removed from the target real-world situation of interest. We thus worry about external validity: what does laboratory evidence tell us about socio-economic reality? Can what is observed in the highly controlled environment of the laboratory be generalised to the more complicated and possibly different real-life situation? Depending on how the two empirical situations (the real life we want to understand, and the laboratory producing accurate measures of the mechanism analysed) are compared, there are different definitions of external validity. We discuss the three most-common definitions, in increasing order of requirements: parallelism, robustness and inference.

One common feature of all of these definitions is that they amount to the comparison between two empirical situations, and are thus empirical in nature. The whole range of empirical methods (laboratory, mobile and online controlled experiments, and field experiments) can be used to 'assess' this external validity, i.e. to check whether the experimental results are sensitive to features that may prevent the results being generalised. The chapter concludes with the existing empirical evidence on this question, and the increasing concern for greater replication, ensuring the credibility of the empirical regularities that are observed in the laboratory.

[8] Some would add to this list the classroom experiments that are increasingly used in economic classes. The seminal example here is the classroom version of the Chamberlain and Smith market experiments, the results of which have, as far as we know, been very well replicated in all instances. Walker (1987) discusses the strengths of classroom experiments in teaching economics; for applications see e.g. Holt and Laury (1997) for the voluntary-contribution game and Holt and Capra (2000) for the prisoners' dilemma. Holt (2006) provides a thorough survey.

9 More Accurate Theory and Better Public Policies: the Contributions of Experimental Economics

Three main reasons regarding the need for experiments were advanced in Chapter 4: testing theory, looking for facts and designing better public policies. These three objectives are well addressed by experiments, as the latter provide a bridge between economic theory and the real world. But experiments are not an exact replica of the real world, and the resulting differences lead to the external-validity issues discussed in Chapter 8. Once these issues have been tackled, the circle can be closed by the following questions: what should we do with the theories once they have been convincingly tested? What do we learn from these results from a public-policy perspective?

The answers to these questions are closely related to the gaps between the behaviour that is expected in economic situations and what is actually observed in the lab. To a certain extent, the answers to these questions are then thus more akin to the concerns of behavioural economics – the literature that brings together economics and psychology in order to offer an empirically more meaningful view of individual decision-making – than to experimental economics per se. As behavioural economics is beyond the scope of this book, we will restrict ourselves here to a few examples aimed at illustrating the ways in which these questions have been answered in the existing literature, with a particular focus on the methodological issues at stake.

The debate between neoclassical economics and behavioural economics is rich and can be considered from a number of different perspectives. One perspective discusses and compares the underlying assumptions of the models, and another is to dismiss the evidence collected in behavioural economics as 'irrational', inconsistent or irrelevant. A third, alternative, perspective is to take the collected body of behavioural economics as an opportunity to make better predictions about the effects of existing policies. This chapter follows this third perspective, and more particularly the review of behavioural economics and public policy proposed by Chetty (2015). The latter distinguishes three main contributions of experimental and behavioural economics. First, the evidence collected by experimental and behavioural economics might help to improve theory, and the objective of improving theory is here to obtain better predictions. Sections 9.1 and 9.2 deal with these implications. Second, even in a neoclassical framework, the experimental method allows the simulation and measurement of the impact of policy interventions. Section 9.3 illustrates this point with a case study of the design of matching markets. Third, the accumulated evidence on behaviour collected from experiments offers new policy tools allowing policymakers to expand the set of policy outcomes. The final Section, 9.4, addresses this point.

9.1 Testing Theory: Drawing General Lessons from (Causal) Experimental Evidence

Experiments that test theory apply a given model in the laboratory and then compare the experimental observations to those predicted by theory. This can apply to the outcomes – the decisions actually taken – but also to the behavioural assumptions of the model. The elicitation procedures presented as case studies in the previous chapters are the standard experimental tools used to test these assumptions. In both cases, there are three broad types of conclusion: (i) accept the theory as it is, (ii) reject it, but also (iii) further qualify the range of its application. Illustration 9.1 provides an example of the latter and shows how experimental behaviour helps to produce a better understanding of the empirical content of size effects in competitive collusion.

A common feature of conclusions (i)–(iii) is that they require a definition of the type of empirical behaviour that should be considered in accordance with or contradictory to a theory. This is not a trivial question following what we learned about theoretical models in Chapter 4. Theoretical models are purposefully wrong as they represent a simple understanding of complex situations. Individual heterogeneity is a classic example of a gap between experimental outcomes and theoretical predictions that is beyond the scope of model testing. Accounting for individual heterogeneity (in beliefs or preferences, for example) is extremely complex in theoretical models, producing probability distributions over different states of the world or wealth distributions, for example, instead of the much more convenient single-outcome prediction in models assuming identical agents. It is obviously unfair to compare the experimental outcomes – from heterogeneous subjects – to the predictions from this model, and it is often worthless generalising the model to include heterogeneity that adds complexity without producing additional insights. The testing of theory thus is a subtle process requiring the definition of which component of the distribution of observed behaviour is expected to inform us about the theory's relevance – see the 'Testing theory' section of Schotter (2015, p. 73) for a detailed discussion.

Using these definitions, there are plenty of examples in which 'economic theory makes strong predictions . . . and is generally quite accurate in predicting behaviour in the laboratory (Levine and Zheng, 2015, p. 43). This stands in contrast to the general impression that the results in experimental economics typically do not provide evidence in favour of rationality or equilibrium reasoning in actual behaviour. This contrast may partly result from the first phase of what Camerer et al. (2003, p. 1216) consider to be the two waves of the recent history of behavioural economics, which 'identified a variety of disparate phenomena that were all anomalous compared to rational choice predictions, but which otherwise had little in common. As a result, early critics of behavioural economics often complained that it was just a laundry list of departures from rational choice'.

The crowding-out effect of monetary incentives, described in Illustration 9.2, is an example of such an anomaly, which with no doubt contradicts one of the most fundamental parts of economic theory: that individuals respond positively to monetary incentives. The aim of this section is to discuss how and to what extent these anomalies

Illustration 9.1

Market size and collusion: 'two are few and four are many'

As discussed in Chapter 4, Illustration 4.2, the ability to collude on competitive markets greatly depends on the size of the market, as measured by the number of firms: collusion is more likely to result with few firms. In an attempt to supply some empirical content to this theoretical insight, Selten (1973) stated that 'four are few and six are many'. Huck et al. (2004) provide an experimental assessment of this claim, based on Cournot competition experiments (combined with a meta-analysis of existing results, which we do not report here). A fixed number N of firms compete in quantity, $q_i \in [0, 100]$, over 25 periods. The linear demand and cost functions are specified in such a way that the one-shot Nash equilibrium is

$$Q^* = \sum_{i=1}^{n} q_i^* = 99N/(N+1)$$

The collusive level of output is $Q^C = 49.5$, and the competitive, zero-profit, quantities are $Q^R = 99$. The between-subject treatment variable is the size of the market, which varies from $N = 2$ to $N = 5$. Overall, six markets are observed for each size, and the exchange rate is adjusted across treatments so that the monetary payoffs remain the same despite the change in equilibrium behaviour . The table below shows the Nash equilibrium, the quantities observed on all markets and those observed in the last periods of the experiment, and for both of these the Pearson correlation coefficient (*corr.*) between the observed quantities and the Nash equilibrium.

		All periods of play			Last periods		
N	Q^*	Q_{1-25}	(St. dev)	$corr._{1-25}$	Q_{17-25}	(St. dev)	$corr._{17-25}$
2	66.00	59.36	(3.76)	0.89	60.44	(7.05)	0.91
3	74.25	73.47	(6.85)	0.99	72.59	(4.53)	0.98
4	79.20	77.26	(7.75)	0.98	80.67	(4.85)	1.02
5	82.50	86.21	(7.11)	1.05	88.43	(8.80)	1.07

As expected, quantities fall with market size. For all sizes, the correlation with the Nash equilibrium is over 80%, and higher at the end of the experiment than at the beginning. The treatment differences are not significant when compared incrementally in size. The difference between markets of size two and four is, however, strongly significant, with the quantities becoming insignificantly different from the Nash equilibrium in the latter case. This supports the main claim of the paper, from an empirical point of view: that two are few and four are many.

contribute to the development of theoretical models. This inductive approach stands in sharp contrast to the classic view of economic theory, which requires the clarification of the aim and applicability of theoretical models. Disagreements in this respect are numerous, and explain the variety of opinions about the empirical content of economic theory (see the controversy in Binmore and Shaked, 2010; Fehr and Schmidt, 2010, for an extreme example of how deep the disagreement can become). We summarise the

Illustration 9.2
The hidden cost of incentives: motivation crowding out

Gneezy and Rustichini (2000) analyse the effect of monetary incentives on performance in two different experiments. In the first, subjects take an IQ test. The between-subject treatment variable is the piece rate offered for right answers: no piece rate at all, a small one (of 0.10 NIS cents, the local currency), a medium one (1 NIS) or a large one (3 NIS). The left-hand side of the table below (from Gneezy and Rustichini, 2000, p. 797, Table 1) lists the average number of right answers (out of 50 questions), along with the standard deviation and median in the sample of subjects.

	Experiment 1: IQ test				Experiment 2: fund raising		
	No payment	10 cents	1 NIS	3 NIS	No payment	1%	10%
Average	28.4	23.1	34.7	34.1	238.6	153.6	219.3
St. dev.	13.9	14.7	8.9	9.4	165.8	143.1	158.1
Median	31.0	26.0	37.0	37.0	200.0	150.0	180.0

Small monetary incentives do not improve performance, and even reduce it. Incentives only improve performance if they are large enough, in which case performance increases by 50%. To check the robustness of these results, a second experiment was run in field conditions: pairs of students participated in a fund-raising campaign (which traditionally takes place in Israel). Again, the between-subject treatment is the piece rate: from no monetary incentives in the baseline, the rate changes to 1% of the total amount collected in the first treatment and 10% in the second. The right-hand side of the table above (from Gneezy and Rustichini, 2000, p. 800, Table 4) shows the amount of money collected by each pair of subjects. Monetary incentives here reduce performance: when there is only intrinsic motivation more money is raised than under performance-based incentives. As above, performance does start to improve again when incentives are high enough – although performance is still worse than with no incentives. These results illustrate robust findings in the literature that extrinsic incentives might crowd out intrinsic motivation (see Frey and Jegen, 2001, for a survey).

main arguments in these discussions in the following section, before moving on to the question of induction itself.

9.1.1 Does Theory Have to Match the Facts?

There are two properties of the usual theoretical approach in economics that are worth keeping in mind when discussing induction. First, theory is a broad term that might refer to two very different objects: normative analysis, which attempts to establish conclusions about what outcomes should be, and positive analysis which aims to say what is expected to happen under a given set of circumstances. The question of behavioural assumptions in normative analysis, and their empirical relevance, will be

discussed in Section 9.4. One important point at this stage is that many anomalies are considered irrelevant from a normative perspective: a failure to behave rationally does not lead us to discard rationality, just as a typo does not lead us to drop grammar.

This does not mean, however, that what makes sense from a normative point of view needs to hold in models that are developed for positive analysis. In this respect, rationality remains a useful guide to the implications of consistent behaviour: theory might be 'wrong' but nonetheless remains 'strong' and provides a useful benchmark for the interpretation of what is actually observed (Schotter, 2006). However, the reluctance of the profession to split up the positive and the normative approaches mainly reflects the second particularity of economic analysis, which has long been hypothetico-deductive in nature: the predictions of outcomes come from consistent assumptions regarding behaviour, and these outcomes are the only thing that matter (see Bardsley et al., 2009, Chapter 4, for a detailed discussion).

This reasoning relies heavily on Milton Friedman's famous exposition of the 'as-if' assumption – the empirical content of behavioural assumptions is irrelevant as long as it leads to reasonable predictions. Since theories are always wrong by construction, meaningful content does not imply that the theory perfectly describes all the details of a given situation, but rather that (i) it contains plausible hypotheses, (ii) it leads to empirically meaningful conclusions and (iii) it is robust to changes in its assumptions (Rubinstein, 2006a). As a result, meaningful content refers to whether the theory is useful or useless, rather than correct or incorrect (Levine and Zheng, 2015). This view does, however, restrict theory to its predictive power. But theory also serves an explanatory purpose (Schotter, 2015): if an outcome is attained and/or changes in response to changes in the environment, we want to know not only that it has happened, but also why it happened. To this end, theoretical models have to be accurate in two dimensions: in terms of how they predict outcomes, of course, but also regarding the driving forces of the behaviour producing these outcomes.

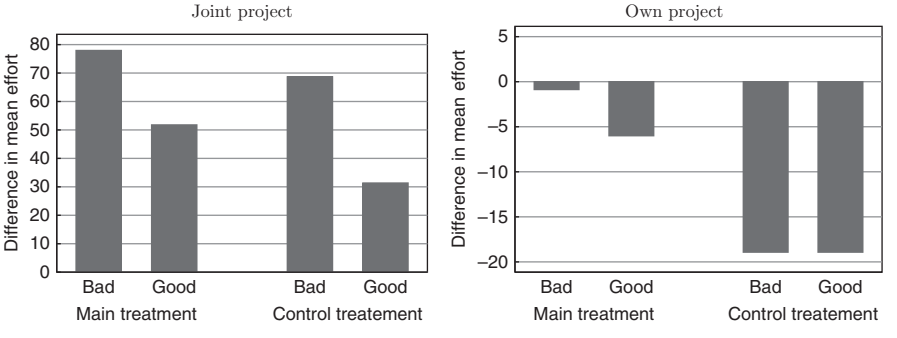
9.1.2 What Kind of Data Deserves Induction?

The previous section suggested that the match of theory and facts is to be understood depending on the objectives we assign to theory. With this statement in mind, we now return to the crowding-out example described in Illustration 9.2. As the relationship between performance and incentives is so central to economic theory, a theory that can explain the crowding-out effect of monetary incentives is clearly required given the existing empirical evidence. The surprising results in the latter suggest that the current state of theory has missed out an important part of the story. As shown in Illustration 9.3, the information conveyed by the choice of performance-based incentives might be an important piece of any missing link when there is crowding-out.

As discussed in Chapter 4, a good theory solves the trade-off between accuracy and parsimony. According to Stigler (1965), theories should then be judged according to

Illustration 9.3
The informational content of incentives: an experimental test

Benabou and Tirole (2003) show that crowding-out can arise in a principal–agent framework due to the informational structure of the relationship. They consider a principal who is better informed about the difficulty of the task than the agent to whom it is delegated. For given agent ability, the harder the task the greater the incentives proposed. This conveys discouraging information to the agent – if the incentives are higher then the goal is less likely to be achieved – resulting in less effort and worse performance with higher incentives. Bremzen et al. (2013) experimentally assess this insight, where a principal receives information about the cost of effort and chooses the fixed wage level as well as an effort-based bonus. The agent then receives a private signal about the cost of effort, and makes two effort decisions: the effort put into the 'joint production' that benefits both the principal (directly) and the agent (through the bonus), and the effort put into her 'own project' that benefits only the agent. The two effort choices allow us to disentangle the incentive effect of the bonus from its informational effect on effort choice. As the cost of effort is the same for both projects, and the bonus only applies to joint production, any change in the effort put into the own project following a change in the bonus identifies the pure informational effect of incentives. In a control treatment, an identical experiment is carried out, except that the principal does not receive any information about the cost of effort. In the experiment, subjects playing the role of the principal are seen to adjust the bonus according to the difficulty of the task, so that incentives are an informative signal about the cost of effort: a high bonus is chosen 80% (32%) of the time when the cost of effort is high (low). The figure below (from Bremzen et al. (2013), p. 62, Figure 3, and p. 63, Figure 4) depicts, for each of the two projects, the difference in the mean levels of effort between bonus and no bonus as a function of the signal received by the agent.



While bonuses do improve performance, the differences in effort put into the own project as a function of effort cost confirm the main behavioural prediction of the model: higher-powered incentives are interpreted as bad news about the cost of effort, which results in lower effort being put into the own project.

three criteria: congruence with reality, generality and tractability. These three elements are to a large extent contradictory, and the choice of the facts that we take as relevant to test the theory is accordingly difficult. In particular, many of the assumptions embedded in theoretical models are not intended to exactly match empirical features, but rather to reduce complex reality to a simple framework. The assumption of rationality arguably plays this role in economic analysis. The assumption that people are selfish, for instance, makes sense not 'because economists believe that people are selfish – we doubt you could find a single economist who would assert that – but rather because in competitive markets it does not matter whether or not people are selfish because they have no opportunity to engage in spiteful or altruistic behaviour'. (Levine and Zheng, 2015, p. 47). As an example, Illustration 9.4 shows that market arbitrage might discipline preference reversals. The need for theoretical foundations of this kind of behaviour thus depends on the circumstances under which the behaviour is analysed. Preference reversal may well be irrelevant if decision-makers are in a market situation – even if it is a real component of individuals' decision-making.

As advanced by Shogren (2006), behavioural failures can rather be seen in this context as market failures – the lack of well-defined incentives is responsible for the appearance of dominated behaviour. 'Institutional design', whose 'goal is to construct an institution or context that can provide the incentives to induce people to act more rational (whether they actually are or not is another question)' (Shogren, 2006, p. 1148), thus appears as an alternative path to induction – which rather aims to adapt theory to behavioural anomalies. These remarks also illustrate that the empirical content of a theory cannot be assessed without defining the range of situations to which it is expected to apply (what Bardsley et al., 2009, Chapter 2, label the 'domain' of theoretical models). This should be understood as restrictive not only for the theoretical model – which should be specified in reference to a particular domain – but also for the experiment. In particular, as rationality is common in theoretical analysis in economics, failure to behave rationally in a given situation cannot be considered as a test of the particular theoretical model being tested in that situation (Sitzia and Sugden, 2011).

We can ask two questions regarding inference when the empirical facts convincingly contradict theory. The first is to decide the range of empirical evidence that converts casual observation into stylised facts that need to be included in the theory. As in the classic example, repeatedly observing that crows are black does not prove the general law that all crows are black. The commonly accepted response to this first question is the reproducibility of experimental findings, and their replication (see Chapter 8, Section 8.4, for a detailed discussion).

The second question concerns the direction in which theory should be amended without throwing the baby out with the bathwater. While it took a great deal of very inspired effort to formalise the meaning of 'rationality' in axiomatic terms (see e.g., Hammond and Zank, 2014), its behavioural content is straightforward and well-defined: this essentially implies the consistency of choice and consequentialism. Departures from rationality, by contrast, can take many different forms, and need not be consistent with each other. As noted by Rabin (1998), the insights from psychological research regarding individual decision-making do not necessarily imply a radical move away

Illustration 9.4
Preference reversal in a market situation

The WTP/WTA discrepancy (see Chapter 5, Illustration 5.1) is an example of a more general behavioural anomaly, preference reversal, that was first documented in betting experiments by Lichtenstein and Slovic (1971). In these experiments, subjects face two different kinds of lottery, a *P-bet* that offers a high probability of a low stake, and a *$-bet* that has a low probability of a very high stake. Preferences over these two are elicited in two different ways: when subjects are asked which lottery they prefer, they tend to favour the *P-bet*; however, when subjects are rather asked which lottery has the higher value, they assign a larger figure to the *$-bet*. This result is robust to more-experienced populations (for instance, casino players in Lichtenstein and Slovic, 1973) and to changes in the experimental design (Grether and Plott, 1979). Chu and Chu (1990) note that this kind of behaviour should not survive market arbitrage: subjects who put a lower value on the *P-bet*, but at the same time have a strict preference for the *$-bet*, can easily be fooled on a market that implements their preferences if they are sold the *P-bet* at their self-reported price, and then offered to exchange it against the *$-bet* which is then bought back at their self-reported price. Two experiments were run to analyse the behavioural consequences of this property. In the first, subjects were asked to choose between six lottery pairs, followed by the elicitation of their minimum selling prices. One of the six lottery pairs was then chosen at random, and that which was selected was played for real. The compensation of the price-elicitation part relies on a Becker–DeGroot–Marshak procedure applied to the same gamble. In the second part of the experiment, subjects face a series of arbitrages by a trader. The arbitrage takes place over a pair of lotteries for which the largest number of preference reversals was observed in the first part of the experiment. Subjects first face the same pair again and are asked both their preference order and the price of both lotteries. In the case of preference reversal between the two decisions, the cheapest lottery is sold to the participant and traded against the dominant one. The dominant lottery is then purchased from the participant at the assigned fair price. The trader in the second part thus instigates a money pump based on the discrepancy between choice and prices. The whole sequence (choice, pricing and arbitrage) is repeated until the preference reversal disappears. The table below (from Chu and Chu (1990), p. 907, Table 2, group B) shows the main results from the experiment.

	No preference reversal	Preference reversal
Before arbitrage	78.2%	21.8%
After 1 transaction	5.4%	16.4%
After 2 transactions	14.6%	1.8%
After 3 transactions	1.8%	0%

The table lists the percentage of subjects (out of 55) who reverse their preferences (second column) or do not (first column), both in the first part of the experiment (the first row) and at each stage of the second part. Preference reversal quickly disappears once the decisions are made in the presence of a market mechanism. Three repetitions suffice to make almost all anomalies disappear.

from the standard assumptions. The extension of preferences to include other-regarding motives or relative rather than absolute outcomes can, for instance, easily be embedded in a rational-choice framework – and arguably even does not contradict its predictions (Levine and Zheng, 2015, p. 46). In this case, experimental evidence allows us to 'consolidate' theory (Binmore, 1999), i.e. to obtain more accurate definitions of the hypotheses and their consequences. By contrast, other aspects of economic psychology, such as judgement biases or anomalies in decisions over time, do contrast with the fundamental assumptions regarding rational decision-making.

The strand of literature that attempts to revise theory to incorporate these features is what Camerer et al. (2003, p. 1216) call the second wave in the recent history of behavioural economics, which 'represents a scientific consolidation [of the criticisms addressed to the first wave]. Precise functions that add one or two free parameters to standard rational theories are being applied to explain important anomalies and make fresh predictions' (see Bruni and Sugden, 2007, for a detailed historical perspective). Rabin (2013, p. 2) recommends this approach to comply with the PEEM principle (portable extensions of existing models), according to which, 'One should (a) extend the existing model by formulating a modification that embeds it as parameter values with the new psychological assumptions as alternative parameter values, and (b) make it portable by defining it across domains using the same independent variables in existing research, or proposing measurable new variables'. The case study in the following section describes an example from game theory.

9.2 *Case study:* Rational Behaviour, Irrational Thinking: *K*-level Models

The guessing game is a good example of an attempt to adapt theory to behavioural regularities. A guessing game involves N individuals, each of whom is asked to choose a number y_i in the interval $[0, 100]$. Once all decisions have been privately made, the average of all of the numbers $\bar{y} = \sum_i y_i/N$ is calculated. The winner of the game is the player whose chosen number is the closest to a share $p \in [0, 1]$ of the group average $\bar{y}$. The winner's payoff is a fixed amount, independent of both their chosen number, y_i, and the game parameter p – whose value is common knowledge before the game starts. In the case of a tie, the payoff is equally split between the winners.

The theoretical prediction of behaviour in this set-up is based on both the common knowledge of rationality and the elimination of iterated dominated choices. To illustrate the reasoning, consider the range of possible choices of y_i in the game with $p = 2/3$. All numbers between 67 ($100 \times p$) and 100 are weakly dominated by 67: as the numbers are bounded by 100, $p \times \bar{y}$ cannot be over 67, so that any choice above this figure is less likely to be further from the target. A rational player should therefore exclude all numbers in the interval [67; 100]. If players also believe that all other players are rational and follow the same reasoning, they should then all believe that nobody will choose a

number in this same interval. As a result, all numbers between $67 \times p$ (≈ 45) and 100 are weakly dominated by $67 \times p$: the target number resulting from the decisions of rational players who believe that others have eliminated numbers over 67 cannot be greater than $67 \times p$. A further round of reasoning suggests that players who know that the others will not choose a number over $67 \times p$ will eliminate all numbers above $67 \times p \times p$, and so on. The process ends when the number 0 is chosen, which is the Nash equilibrium of the game when all players choose 0; they all choose the number that is exactly equal to $p \times$ the target. Formally speaking, $y_i = 0$, for all participants i is the unique pure-strategy Nash equilibrium, obtained by the iterative elimination of weakly dominated strategies.

This theoretical prediction holds for any value of p, as long as $0 \leq p < 1$. If $p = 1$, the guessing game corresponds to the beauty contest in Keynes (1936), in which the winners are those who correctly guess the average of all of the chosen numbers. In this case, the game reduces to a coordination problem: all players have to coordinate on the same number – and every such number is a Nash equilibrium of the game. The reasoning behind this result is both convincing and elegant, but relies on very demanding assumptions about players' 'depth of reasoning' – their reliance on high-order beliefs, involving not only what they think others will do, but also what they think others think they themselves will do, etc.

9.2.1 The Empirical Depth of Reasoning

Nagel (1995) designed an experiment aimed at eliciting subjects' depth of reasoning based on their behaviour in a guessing game. The experiment is based on sessions with 15–18 subjects playing the game together four times, with full feedback from one round to the next. The feedback includes all the numbers chosen by each of the N participants $y_i, i = 1, \ldots, N$, the mean $\bar{y}$, the product $p\bar{y}$ and the winning number. Two main treatments are considered, with different levels of p: $p = 1/2, 2/3$.[1]

The distribution of behaviour observed for each treatment in the first round is displayed in Figure 9.1. In neither of the two treatments did any subject choose the predicted number $y_i = 0$, and only 6% of the chosen numbers fell below 10. While this is not compatible with the theoretical prediction, purely naive choices are just as uncommon: few subjects only make the weakly dominated choices comprised between $100 \times p$ and 100. Moreover, subjects do not pick their numbers at random: subjects choose higher numbers in treatment $p = 2/3$ than in treatment $p = 1/2$. Bosch-Domenech et al. (2002) replicate these patterns of behaviour in the field, with over 7,500 volunteers recruited via newspapers or magazines (the *Financial Times* in the United Kingdom, *Expansión* in Spain and *Spektrum der Wissenschaft* in Germany). They find spikes in the distribution at 33.33, 22.22 and 0. In sum, the observed behaviour is consistent neither with fully rational players, nor with naive or random behaviour. The

[1] A third value, $p = 4/3$, is used as a control – since p is greater than 1, there are two Nash equilibria using the same reasoning as before: 0 and 100. The results from this treatment are not discussed here.

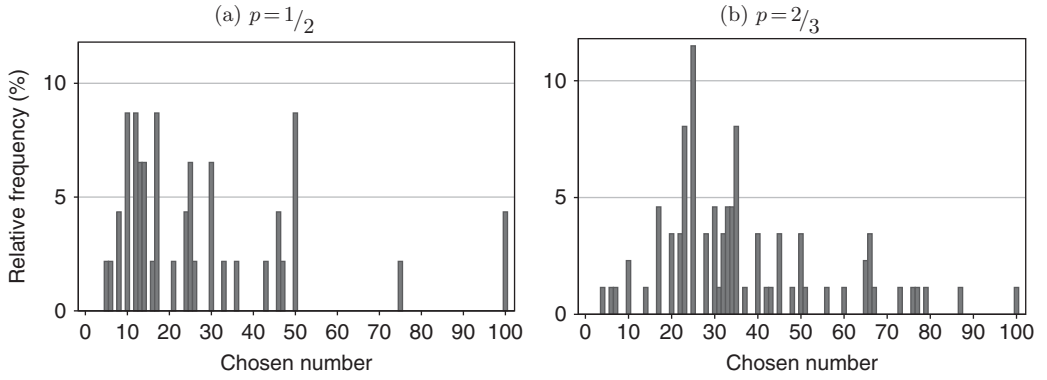


Figure 9.1 The chosen numbers in the Nagel (1995) guessing games
Note.
Source: Nagel (1995, p. 1316, Figure 1).

reasoning behind behaviour rather seems to lie somewhere in between those two polar cases.

To shed light on observed choice behaviour, Nagel (1995) developed a simple model in which players differ in their degree of sophistication – sharing some common features with Stahl and Wilson (1995). The starting point is a particular specification of the behaviour of degree-0 players, who do not think strategically at all. A common choice is to assume that these players pick 50, which corresponds to the average value if everybody else chooses at random – this choice is supported by the empirical evidence provided by Burchardi and Penczynski (2014). The best reply to this behaviour is to choose $50 \times p$: this is the decision of a rational player who thinks all others are of degree 0. As this involves a higher degree of strategic reasoning, this is called degree-1 strategic behaviour. Similarly, the best reply to degree-1 strategic behaviour is to chose $50 \times p^2$. In general, a player with strategic behaviour of degree h chooses $50 \times p^h$. This simple model produces a number of modes in the distribution of observed behaviour: these modes reveal the degree of player sophistication, with smaller values of the mode corresponding to a greater number of iterations in the reasoning leading to this action.

The experimental results discussed above are in line with this choice model: the distribution of answers is multimodal, with modes mostly around the values of 50, $50 \times p$ and $50 \times p^2$. Another critical aspect of the data is that about 25–30% of subjects choose a number that turns out to be optimal given the observed distribution of answers. In the treatment with $p = 2/3$, for example, 0 is the optimal choice only under common knowledge of rationality. This no longer holds if players exhibit varying degrees of strategic behaviour. Given the observed behaviour in the subject population, the optimal choice in the experiment is about 25, which corresponds to strategic behaviour of degree 2. In this framework, the gap between optimal behaviour and bounded rationality arises not because of rationality failure, but rather because the decision-maker does not believe that other players will choose an equilibrium strategy.

This simple model also allows us to take into account changes over time due to feedback: if information about others' past behaviour is provided, the number should change accordingly and will differ from 50. The changes over time in Figure 9.2 support this finding. For each of the two specifications of the guessing game, the top panel shows the distribution of numbers chosen in the second game (after subjects received feedback) as a function of the first-period decisions, the middle panel the third-period decisions as a function of second-period decisions, and the bottom panel fourth-period decisions as a function of the third-period decisions. Rather than a change in levels of strategic sophistication over time, the figures show that subjects tend to adapt their beliefs about the likely behaviour of others based on previous play. Subjects do understand the reasoning required to win this game, but to some extent fail to anticipate others' behaviour when choosing their own number. This produces convergence towards the Nash equilibrium with more repetitions of the game.

9.2.2 The Level-k Specification of Strategic Reasoning

The results in the guessing game point to limitations in individuals' abilities to think about others' behaviour in the depth presumed by the rational model of decision-making. Similar evidence has been found on experimental asset markets (Akiyama et al., 2017) where strategic uncertainty explains at most 70% of the observed deviations from the fundamental values. Many attempts to incorporate more flexible kinds of strategic thinking have been proposed in the literature – e.g. equilibrium plus noise, k-rationalisability and finitely iterated strict dominance. We herein focus on one of the most popular, the level-k specification that generalises the behavioural model developed in the previous section. This class of models assumes that individuals form beliefs over their opponents' actions in discrete steps. Each player in a game is characterised by a 'type' describing their degree of sophistication about what they believe others will do:

- A level-0 (or L_0) player does not play strategically. The actions of L_0 players may be random, the best response to the rules of the game (ignoring opponents' behaviour), or use a focal point.
- A level-1 (or L_1) player plays strategically and best-responds to the beliefs of all L_0 players. The actions of L_1 players reflect their beliefs about the distribution of actions of L_0 players.
- A level-2 (or L_2) player plays strategically and best-responds to the beliefs of all players that are L_1. The actions of L_2 players reflect their beliefs about the distribution of actions of L_1 players.
- ...
- A level-k (or L_k) player plays strategically and best-responds to the beliefs of all players that are L_{k-1}. The actions of L_k players reflect their beliefs about the distribution of actions of L_{k-1} players.

A core assumption is that all players L_k, with $k > 0$, are rational, in the sense that they best-respond to their beliefs, although these beliefs differ from those in standard game

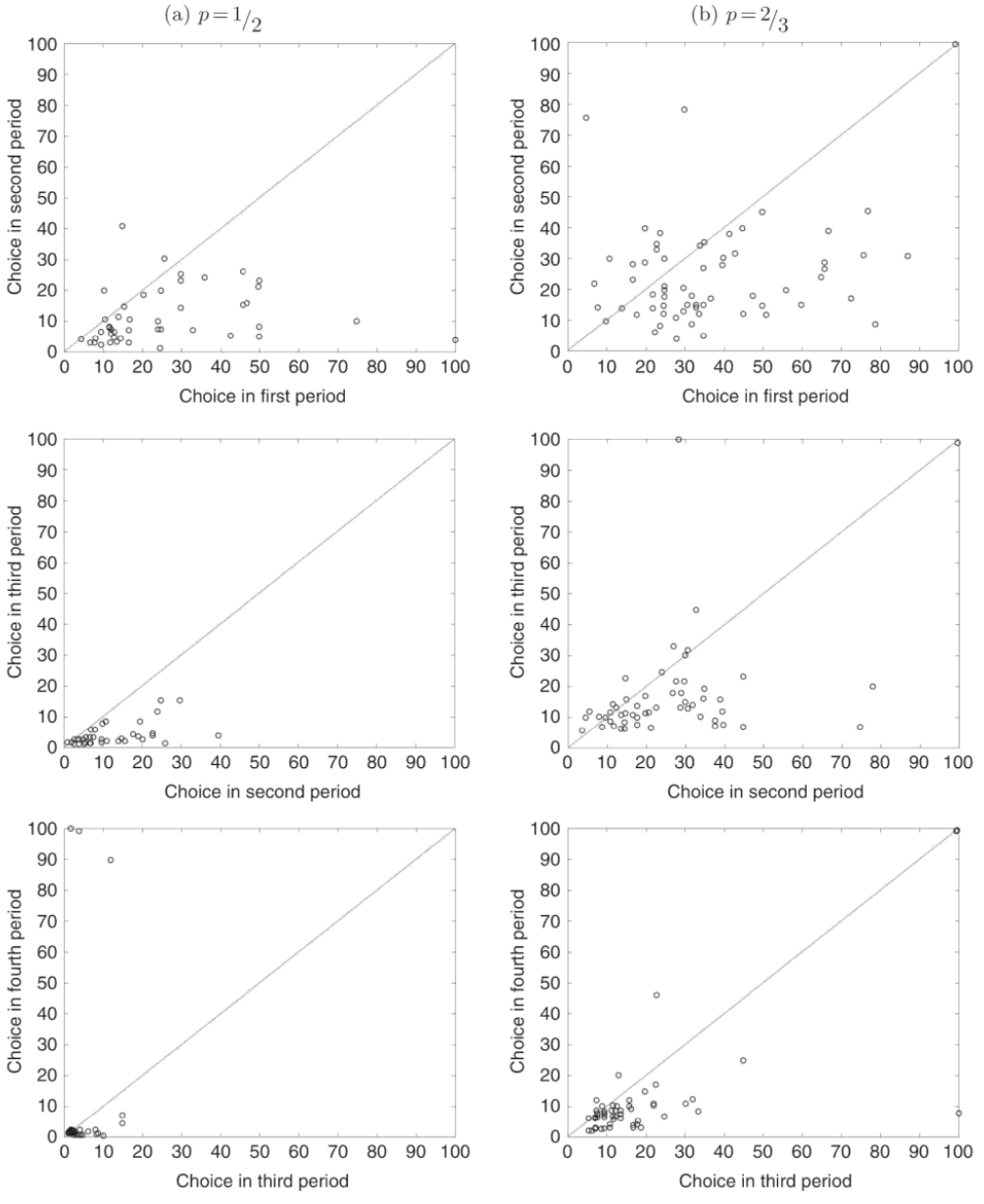


Figure 9.2 The distribution of behaviour over time in the guessing game

Note. For each of the specifications of the guessing game indicated in the column head, each graph depicts the distribution of numbers chosen in the second, third and fourth rounds as a function of the number chosen the previous period. The 45° line shows the expected behaviour of subjects who continue to make the same decision.

Source: Nagel (1995, p. 1319, Figures 3 and 4).

Illustration 9.5
The market-entry game

In a market-entry game, N players (or entrants) simultaneously decide whether to enter a market with limited demand $y_d < N$ or to stay out. If y_d or fewer players enter, all entrants earn a positive profit: if more than y_d players enter the market, all entrants lose money. Staying out yields a profit of zero. Standard game theory predicts a unique mixed-strategy equilibrium in which the probability of entering makes all players indifferent between the two possible actions. A large body of experimental evidence (Camerer and Loewenstein, 2003) has shown that actual entry rates are remarkably similar to that predicted in the symmetric equilibrium. Moreover, the number of entrants rises with the capacity constraint y_d, even though players have no way of deliberately coordinating their choices. Observed behaviour, however, differs from the equilibrium prediction for extreme values of y_d: for low values subjects tend to over-enter the market, and for high values they under-enter. As shown in Camerer et al. (2004) and Crawford and Iriberri (2007a), these results can be explained by strategic thinking, and especially the 'magic of tacit coordination' (Kahneman, 1988) achieved in market-entry games. As players are heterogeneous, higher-level players are able to predict the decisions of less sophisticated players. Strategic thinking, because of its heterogeneous iterative mental structure and the replication of lower types in the simulations made by higher types, explains why coordination can occur in a simultaneous one-shot game played by independent players.

theory. In one-shot games, for example, players no longer believe that others will play an equilibrium strategy, but rather that they have a model of thinking that is simpler than their own. Illustration 9.5 describes an application to a game that is strategically similar to a guessing game: a market-entry game.

The level-k model is fairly simple to solve thanks to the assumption that players best-respond to the behaviour at the level just below their own. This is the main difference from the cognitive-hierarchy model, presented in Focus 9.1, which extends the framework by assuming that players respond to a mixture of all lower-level players.

In practice, the actions of level-k players are determined recursively from the knowledge of the behaviour of L_{k-1} players. This recursive solution starts with the behaviour assumed for L_0 players. There are then two crucial elements in this model: the actions of L_0 players and the distribution of each type. It is worth noting that there may not be any L_0 players in the distribution: what matters is that the L_0 actions are used by the L_1 to form their beliefs about players' behaviour in the game, not that the L_0 actions are effectively undertaken. Based on this structure, level-k models treat deviations from equilibrium as deterministic. This is the main difference from one of the most popular alternatives, the QRE model presented in Focus 9.2, in which these deviations are treated as noise or responses to noise.

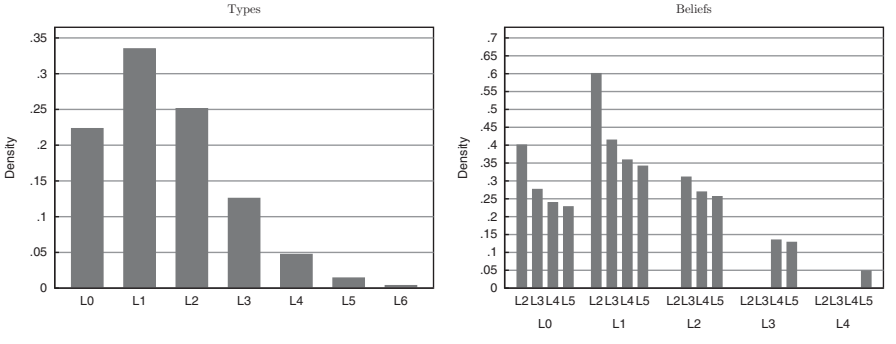
A number of studies have underlined the use of level-k reasoning to rationalise behaviour in guessing games. For example, Costa-Gomes and Crawford (2006) elicit

Focus 9.1
The cognitive-hierarchy model

The cognitive-hierarchy model (CHM) (Camerer et al., 2004) features players who believe that they each understand the game better than all other players. Decision-making is assumed to be based on iterative decision rules, involving k steps of thinking. In the guessing game, for example, L_2 players believe the population of other players to be a mixture of L_0 players, guessing 50 on average, and L_1 players, guessing $\frac{2}{3} \times 50$. The frequency distribution $g(k)$ of level-k players is assumed to be Poisson. For $k \geq 1$, an L_k player normalises the actual frequencies to form beliefs over all other (lower) types $h = 0, \ldots, k-1$ according to

$$g_k(h) = \frac{g(h)}{\sum_{j=0}^{k-1} g(j)}$$

In this framework, beliefs are increasingly accurate as k rises. Moreover, the benefits from deeper thinking are diminishing: the difference between successive levels disappears as k grows. Camerer et al. (2004) present estimates of the Poisson parameter in 24 guessing games with different values of p. They find a reasonable value for this parameter of 1.61 – with some heterogeneity across populations: the value being, for instance, much higher for professional stock-market portfolio managers, Caltech students and game theorists. In entry games (see Illustration 9.5), the value is typically between 1 and 2. Camerer et al. (2004) suggest that a value of 1.5 could produce reliable predictions for many other games. To illustrate the functioning of the model, the figure below displays simulations of player-type distributions (on the left) and belief distributions (on the right) for players L_2 to L_5 under this specification.



The comparison of the distributions of L_4 and L_5 players illustrates the diminishing benefits from thinking. The marginal gains from thinking are higher for L_2 to L_4 players and lower for L_5 players. The values of the Poisson parameter distribution also differ within subjects across games.

Illustration 9.6
Strategic thinking in the centipede game

As discussed in Chapter 1, Section 1.3.2, the experimental evidence on the centipede game stands in sharp contrast to the theoretical predictions from backward induction. Kawagoe and Takizawa (2012) use level-k and cognitive-hierarchy models to analyse behaviour in centipede-game experiments, building on the original design of McKelvey and Palfrey (1992, 1998, with increasing amounts of money) and Fey et al. (1996, with a constant amount of money). The two strategic-thinking models are tested with both parametric and non-parametric distributions of types. Different kinds of L_0 players are also considered: pure randomisers or altruistic players choosing P at all nodes. The results show that a cognitive-hierarchy model based on a Poisson distribution and with L_0 players choosing at random is the best specification for centipede games with increasing amounts of money. For games with a constant amount of money, where choices are closer to the sub-game perfect Nash equilibrium, however, the QRE model including a fraction of altruistic players (McKelvey and Palfrey, 1992, 1995) performs best. Alternative explanations that have been considered in the literature are adaptative learning (Rapoport et al., 2003) and other-regarding preferences (Dufwenberg and Kirchsteiger, 2004).

initial responses to 16 different games (the subjects are rematched with new partners at each period and receive no feedback). In this experiment, subjects face a large strategy space and both the target of the guessing game and the range in which guesses are made vary independently across players and games. The results show that most subjects try to maximise payoffs, but that their representation of others' likely behaviour leads to systematic deviations from equilibrium: other individuals are mainly thought to be L_1, L_2 and L_3 types. The following section shows how the model helps to understand behaviour in auctions – see also Illustration 9.6 for the centipede game, and Crawford et al. (2013) for a survey.

9.2.3 Level-k Reasoning and Behaviour in Auctions

A robust stylised fact in empirical auctions is that subjects tend to overbid – see e.g. Kagel and Levin (1986); Goeree et al. (2002). In common-value auctions, this phenomenon is the well-known winner's curse – bidders fail to account for the common-value nature of the auction – discussed in Illustration 8.1 in Chapter 8. In private-value auctions, overbidding is usually explained by risk aversion, the joy of winning or non-linear probability weighting. Crawford and Iriberri (2007b) appeal to level-k strategic thinking to provide a unified explanation ('level-k auction theory') of overbidding in both independent-private-value and common-value auctions.

As always in level-k models, the difficulty is to define the behaviour of the L_0s. Crawford and Iriberri (2007b) consider two specifications:

- Random: L_0 players bid randomly from a uniform distribution over the feasible range. L_1 and L_2 types faced in this specification base their actions on random L_0 players.

Focus 9.2

An alternative theoretical model of strategic thinking: quantal-response equilibrium

Probabilistic-choice models (Goeree et al., 2005) have been widely used in the analysis of experimental data to include noise in decisions, reflecting, for instance, the errors individuals make (Harless and Camerer, 1994; Hey and Orme, 1994). The quantal-response equilibrium (QRE) model, proposed by McKelvey and Palfrey (1995), incorporates these noise elements into the theoretical analysis of behaviour in games. The core assumption is that players use *noisy* best responses. The distribution of actions under this noise is described by a quantal-response function that replaces the usual best-response functions – according to which players choose the best response to their beliefs about the other players' strategies with probability 1. Under QRE, the best strategy is not necessarily chosen for sure, although strategies with higher expected payoffs are more likely to be chosen. In this sense, the QRE model assumes a form of bounded rationality. Importantly, this framework abandons neither equilibrium reasoning nor rational expectations. An equilibrium is reached when the beliefs of each player are consistent with the (noisy) actions of others – the QRE is a fixed point in the space of the distribution of actions, with each player's distribution being the noisy best response to others'. Compared to models representing choices as 'equilibrium plus noise,' under QRE each player responds to the noise in others' decisions. In empirical applications, it is commonly assumed that players' actions follow a logistic distribution with a precision parameter λ. Let a_{-i} be the strategy profile of players other than i, $\pi(a_i, a_{-i})$ the gain for i corresponding to the strategy profile (a_i, a_{-i}) and $Pr[a_{-i}]$ the probability that other players choose the profile a_{-i}. The probability that player i plays action a_i in the set of all possible actions A_i is then given by a logit formula:

$$Pr[a_i] = \frac{e^{\lambda \sum_{a_{-i} \in A_{-i}} Pr[a_{-i}]\pi(a_i, a_{-i})}}{\sum_{a_j \in A_i} e^{\lambda \sum_{a_{-i} \in A_{-i}} Pr[a_{-i}]\pi(a_j, a_{-i})}}$$

When the precision, λ, is 0, players choose their actions at random and the choice probabilities are uniform over all possible actions. When precision is perfect ($\lambda \to \infty$), the probability of choosing the best response is 1, and the QRE equilibrium corresponds to the Nash equilibrium. The major difficulty of QRE is that it can rationalise virtually any distribution of actions in normal-form games based on an appropriate specification of the distribution of the noise (Haile et al., 2008). As a consequence, the QRE model cannot be falsified in its general form. From the point of view of the analysis of experimental data, however, the *logit quantal response equilibrium* has been shown to reproduce the empirically observed distributions of actions (Goeree et al., 2005).

- Truthful: L_0 players bid their expected value conditional on their own private information. L_1 and L_2 types in this specification base their actions on truthful L_0 players.

There are two key elements in the description of equilibrium strategies in auction theory. First, the bid is chosen trading off the cost of paying a higher price and the benefit of a higher winning probability – the second-price auction eliminates

this trade-off. Second, when the value is imperfectly known, which is typically the case in common-value auctions, the value of the good conditional on winning should be adjusted according to the informational content of this event. Depending on the nature of the good (common-value versus independent-private-value auctions) and the auction mechanism (first-price versus second-price auctions), the optimal bid of L_k players takes into account either value adjustment or the cost–benefit trade-off or both.

For instance, as random L_0 players bid randomly over the feasible range, their bids contain no information about the value of the good in the auction. L_1 players, who believe all other players to be random L_0, therefore choose their bid based solely on their own signal. Their bid is thus an increasing function of their private signal. The L_2 players in this environment take this behaviour into account in determining their optimal bids. The level-k model implies that random L_2 players adjust their beliefs about the expected value of the good using the information revealed by the event of winning the auction – they know that if they win the auction, others' private signals about the good's value are likely to be lower than their own, so that the latter overestimate the good's value. In this setting, value adjustment then makes bids strategic substitutes: this means that if the L_1 players overbid, then the L_2 players will underbid. Crawford and Iriberri (2007b) provide more detailed results depending on the type of auction, and derive the equilibrium with truthful level-k players.

9.2.4 Observing the Empirical Distribution of Types

The predictive power of level-k models relies on the distribution of types: once the model has been solved recursively, the distribution of behaviour can be predicted only when the fraction of each type in the population is known. The empirical identification of this distribution is rendered more difficult by the hierarchical nature of the model: for any L_k, with $k > 0$, both beliefs about others and level of sophistication are correlated in the data.

A number of different approaches have been used in the literature to measure the distribution of cognitive levels. In Agranov et al. (2012), the correlation between beliefs about others and cognitive levels is broken by having subjects play a guessing game with $p = 2/3$ against a set of computers that play uniformly over the support – a strategy that is common knowledge. This corresponds to L_0 players. As a consequence, in this COMPUTER treatment, the numbers elicited reveal the ability to behave like a $L_k, k > 0$ player conditional on the predetermined L_0 behaviour. Two additional treatments are considered: a CONTROL treatment in which subjects (undergraduate students) play against each other, and a GRADUATE treatment in which subjects play against graduate students who are trained in playing guessing games. The robustness of the method is assessed using the strategy method to obtain subjects' choices in the guessing game when opponents are different mixes of 'graduate' opponents and computers.

The experimental results are summarised in Table 9.1. The level classification in the three treatments is the same as in Nagel (1995), whose results also appear in the table

Table 9.1 Level classification in the control, graduate and computer treatments

	CONTROL	GRADUATE	COMPUTER	Nagel's data
L_0	8%	10%	9%	7.5%
L_1	25%	20%	49%	26.0%
L_2	18%	20%	–	24.0%
L_3	8%	5%	–	2.0%
L_∞	0%	10%		
Not classified	41%	35%	42%	40.5%

Note. For each of the column treatments, the table shows the share of subjects who are classified in the row level. The classification follows the method used in Nagel (1995), whose data appears in the last column.
Source: Agranov et al. (2012, pp. 456–7, Tables 2 and 4).

for the sake of comparison. Relative to the CONTROL treatment, the GRADUATE treatment produces an upward shift in the distribution of observed cognitive levels. On the contrary, the distribution of observed cognitive levels in the COMPUTER treatment shifts towards lower cognitive levels compared to the CONTROL treatment. The results from the COMPUTER treatment also show that about half of the subjects can reason at a level of at least 1.

Arad and Rubinstein (2012) design a game (the '11–20' game) that is specifically intended to measure the distribution of types. In this game, L_0's action is unambiguous and the identification of types based on the observed decisions is unique, based on L_0. The game is as follows:

'*You and another player are playing a game in which each player requests an amount of money. The amount must be (an integer) between 11 and 20 shekels. Each player will receive the amount they request. A player will receive an additional amount of 20 shekels if they ask for exactly one shekel less than the other player. What amount of money would you request?*'

A player who ignores strategic thinking will simply request the highest certain amount (20 shekels) and give up any additional payment resulting from strategic thinking. For strategic players, on the contrary, the description of the game actually corresponds to their best-response function. An L_1 player will best-respond to the L_0 action and ask for 19. An L_2 player thus asks for 18, and so on. Moreover, the game is robust to the specification of L_0 behaviour: for an L_1 player, answering 19 remains the best response as long as 20 is the most probable strategy.[2]

Table 9.2 shows the distribution of choices in the experiment, along with the distribution that would result in a symmetric (mixed-strategy) Nash equilibrium, assuming that players maximise expected monetary payoffs. A total of 74% of subjects choose numbers between 17 and 19, corresponding to L_3, L_2 and L_1 types. This is much higher than the 45% predicted in the Nash equilibrium. The share of L_0-type subjects (6%) is much lower than in other experiments. The estimation of a cognitive hierarchy model

[2] Arad and Rubinstein (2012) also show that the game is robust to a wide range of beliefs about others' types, as well as to social preferences.

Table 9.2 The distribution of behaviour in the 11–20 Game

Action	11	12	13	14	15	16	17	18	19	20
Equilibrium (%)					25	25	20	15	10	5
Results (%)	4	0	3	6	1	6	32	30	12	6

Note. For each number that can possibly be chosen in the columns, the first row shows the observed distribution in the experiment, and the second the distribution predicted in the symmetric Nash equilibrium.
Source: Arad and Rubinstein (2012, p. 3565, Table 1).

yields an estimated parameter for the Poisson distribution of 2.36, reflecting the higher level of strategic thinking in the experiment.

Beyond these two examples, a flourishing literature has tried to elicit the distribution of types in experimental games. To name a few examples, Costa-Gomes and Crawford (2006) monitor the search for hidden payoff information by subject in games of various types, and evaluate which level of cognitive reasoning is compatible with various search patterns; Costa-Gomes and Weizsäcker (2008) elicit subjects' beliefs about others' decisions using a quadratic scoring rule in a series of 14 asymmetric two-person 3 $\times$ 3 games; and Burchardi and Penczynski (2014) pair subjects in dyads with common decisions and common objectives and study the transmission of information between partners.

9.3 Test-Bedding Public Policies in the Laboratory: The Example of Matching Markets

Economics is useful in the design of public policies in that it provides an understanding of the kind of behaviour, interactions and hence outcomes that arise from a given set of rules. It is important for the design of economic policy to understand, for instance, how competition works and how firms decide on their prices as this provides guidelines on how markets should be organised and what kind of practice should be curtailed to promote competition – and provides insights into whether or not competition is warranted. Experiments help with policy implications by assessing the empirical content of the assumptions used in the theories. The contribution of experimental economics to the design of public policies is not, however, restricted to those experiments that test theory. When the situation is too complicated for theory, experimental facts can serve as a substitute for convincing theoretical predictions, in order to test-bed any intended changes in the decision environment.

Laboratory experiments have been applied to a wide variety of policy questions, such as competition policy, auctions, pollution-permits markets and food safety (see Normann and Ricciuti, 2009; Noussair and van Soest, 2014, for a survey). To illustrate, we here focus on the example of matching markets. We describe first why matching markets are important from a policy perspective, and then the experimental environment used in the literature to assess their allocation properties – successfully enough for one of the main

contributors to be awarded the Nobel Prize for his work on this topic (see Chapter 1, Section 1.1 for more details).

9.3.1 The Policy Challenges of Two-Sided Matching Markets

In two-sided matching markets each agent on either side of the market has preferences over the possible matches with all agents on the other side. Typical examples of such markets are admissions to universities, the allocation of pupils to schools, the allocation of public housing (and marriage, notwithstanding that the idea of a market is somewhat less natural in this context). A famous practical example is the National Resident Matching Program in the US, through which doctors and residency programs are matched.

In all of these examples, there is a market matching problem. The final allocation consists of matches, i.e. one-to-one mappings between the agents on either side of the market (in the case of schools, for example, the agents would be considered as seats in the school). When the market is decentralised, matching takes a considerable time: agents need to meet to decide whether to match or move on to another option; they want to wait for enough offers to be received and considered before concluding on a match. This leads to congestion on the market, which economic agents typically deal with by acting well in advance – Kagel and Roth (2000) cite the example of medical students in the UK in the 1960s, who were hired a year and half prior to graduation.

The natural answer to this unravelling, where there is always an incentive to match before everyone else, is to introduce centralised clearing houses. Here agents on both sides report their preferences and a matching technology is used to decide the final allocation. It is, of course, generally not the case that all preferences are jointly consistent: whenever there is conflict in agents' preferences, choices need to be made in comparing and aggregating preferences to yield the final allocation. A variety of such choices are contained in different matching mechanisms, the properties of which are the subject of a large theoretical literature (see Roth and Sotomayor, 1992, for an early survey). One critical property of matching mechanisms is whether they produce stable matches, such that (i) no agent prefers being single to being matched with their assigned partner, and (ii) no two agents prefer each other to the partner they have been allocated. Kagel and Roth (2000) provide an experimental analysis of both questions: how clearing houses solve unravelling in decentralised markets, and the stability properties of different matching mechanisms.

9.3.2 Design of the Experiment

The experimental markets consist of 12 subjects, half of whom are assigned the role of a firm and the other half workers. Productivity differs within both groups of firms and workers: half (three) in each group are low-productivity agents and the other three are high-productivity. Subjects are randomly assigned to experimental markets, groups and productivity types.

Performance-based incentives depend on the final match. Remaining unmatched produces no payment. Being matched with a high-productivity agent leads to a payment three times larger than a match with a low-productivity agent. All subjects on one side of the market therefore prefer being matched with a high-productivity rather than a low-productivity partner. However, there is variability within each productivity subgroup, with each match producing a different payment: as a result, individuals' preferences over their potential partners differ.

The experiment features a large number of sequential matching markets. Each matching market is divided into three periods, during which firms make offers to workers, who can either accept or refuse them. Once a worker has accepted an offer, the two matched agents exit the market. Matching is possible in any period, but a cost is imposed on early matches – those achieved before the third period. Conditional on the productivity of the partner, later matches are better. The cost is, however, specified so that it is more desirable to match early with a high-productivity partner than to match with a low-productivity partner in the last period.

The experimental session is split up into two parts: the first with 10 decentralised markets organised as described above. In the second part, subjects are informed that the third period of each of the remaining 15 matching markets will now follow the rules of a centralised market: unmatched agents are asked to submit a rank-order list of their possible matches and the allocation is decided by a matching algorithm. Two centralised allocation mechanisms are considered in this third period: the Newcastle algorithm and the Gale–Shapley algorithm, implemented as between-subject treatment variables.

The Newcastle algorithm is a priority algorithm. After each agent has ranked the alternatives, the algorithm matches agents based on their mutual ranking by defining a level of priority. Priority is defined as the product of mutual rankings. For example, if a worker and a firm rank each other first, then this matching will be of priority 1. However, if they mutually rank each other 4, the matching will be of rank 16. The algorithm then matches agents sequentially by increasing order of priority – couples of priority 1 first, then 2, then 3 and so on. To give an example, consider six firms (a^1, a^2, a^3, a^4, a^5 and a^6) that are to be matched with six workers (b^1, b^2, b^3, b^4, b^5 and b^6). Assume each agent has ranked the potential partners as on the left-hand side of Table 9.3. In each cell, the first value is the rank given by the firm (in the row) to the worker (in the column), and the second value is the rank given by the worker to the firm. The right-hand side of the table displays the Newcastle algorithm match score. The final matching is by ranks. The pair (a^1, b^1) has rank 1 and is matched first. There are a number of matches of rank 2: (a^3, b^4), (a^4, b^3), (a^3, b^3) and (a^4, b^4). Whatever the choice, the pair with rank 3 is (a^2, b^2). The last two pairs are (a^5, b^6), with rank 6, and (a^6, b^5) with rank 12. This last match is not stable as worker 5 and firm 5 would both prefer to be matched together and have an incentive to match together before the implementation of the algorithm.

The Gale–Shapley algorithm (also known as deferred acceptance (DA)) instead considers each worker in turn, to whom a firm is attributed. If the firm is still available, the match occurs. If the firm has already been matched, but prefers the new proposed

Table 9.3 The Newcastle algorithm: a fictional example

	Self-reported rank of potential partners						Resulting scores of all matches					
	b^1	b^2	b^3	b^4	b^5	b^6	b^1	b^2	b^3	b^4	b^5	b^6
a^1	(1,1)	(2,2)	(3,3)	(4,4)	(1,5)	(3,6)	1	4	9	16	5	18
a^2	(2,1)	(1,3)	(4,2)	(3,4)	(2,5)	(2,6)	2	3	8	12	10	12
a^3	(3,3)	(3,4)	(2,1)	(1,2)	(4,5)	(4,6)	9	12	2	2	20	24
a^4	(4,3)	(4,4)	(1,2)	(2,1)	(5,5)	(5,6)	12	16	2	2	25	30
a^5	(5,1)	(5,2)	(5,4)	(5,5)	(3,3)	(1,6)	5	10	20	25	9	6
a^6	(6,3)	(6,1)	(6,4)	(6,5)	(6,2)	(6,6)	18	6	24	30	12	36

Note. The table shows a fictional example of the allocation produced by the Newcastle algorithm. The left-hand side shows the rank given by firms (in rows) to workers and by workers to firms. The right-hand side displays the scores according to which the allocation is decided.

worker, then a new match takes place. Otherwise, the firm rejects the worker and a new firm is proposed to the worker until all matches are made. Based on the preferences displayed on the left-hand side of Table 9.3, suppose workers b^1, b^2, b^3, b^4 and b^6 are temporarily assigned to the firms a^1, a^2, a^3, a^4 and a^5. The algorithm then offers to match worker b^5 with firm b^1: the firm declines since the current match is better than the new one. The algorithm moves to a second choice for worker b^5, say firm a^2, which also declines. Finally, the algorithm offers worker b^5 to be matched with firm a^5. This firm prefers the new offer, which is accepted, and breaks the pre-existing match with worker b^6. This worker b^6 thus needs to find a new match. Say firms a^5, a^2, a^1, and a^4 are proposed in sequence. They all decline but the last firm, a^6, which is currently unmatched, accepts the offer. The process is over, and the final allocation is (a^1, b^1), (a^2, b^2), (a^3, b^3), (a^4, b^4), (a^5, b^5) and (a^6, b^6). In contrast to the outcome from the Newcastle algorithm, this matching is stable.

9.3.3 The Results of the Experiment

Figure 9.3 displays the main outcome from the experiment regarding early matches, i.e. the number of matches observed in the first (Figure a) and the second (Figure b) periods – a measure of unravelling.[3] In both figures, the first 10 periods of play feature only decentralised markets. The remaining 15 are played with either the Newcastle or the DA algorithm in the last market period.

The experimental market produces comparable behaviour to that previously observed in decentralised two-sided matching markets. Subjects quickly adjust their strategies to the institutional environment (very little change occurs from the first five markets to the next five) and significant unravelling is observed. Despite the loss incurred, a significant number of subjects decide to match early on: almost half of the six matches that should occur on each market come about before the third period. As

[3] Kagel and Roth (2000) use the level of the cost imposed on early matches as an additional treatment variable. We here report only the results from conditions with high mismatch costs.

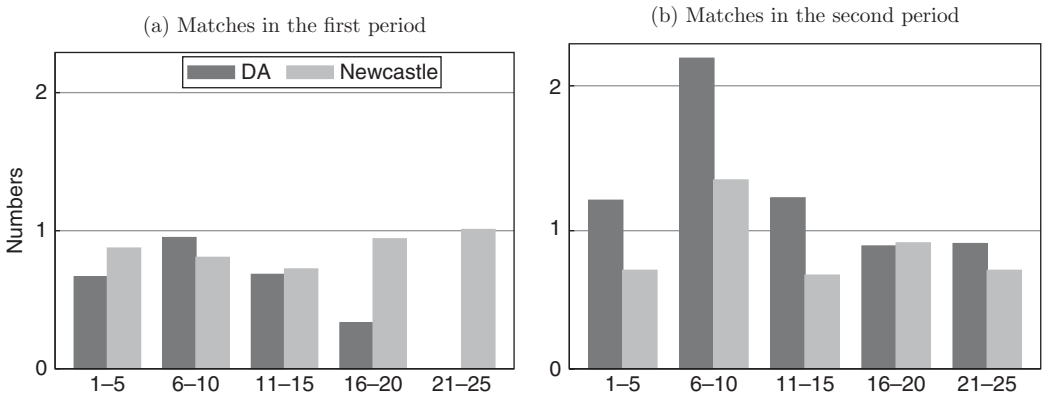


Figure 9.3 Early matches in the Kagel and Roth (2000) experiment

Note. The figure shows the average number of early matches observed in each sequence of five markets. After the first 10 matching markets, the third market period implements either the Newcastle or the DA matching algorithm.

Source: Kagel and Roth (2000, p. 214, Figure 1, bottom panel).

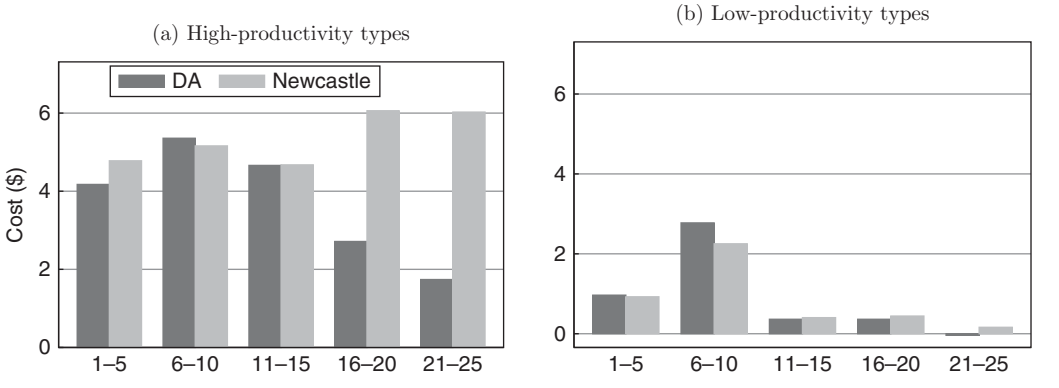


Figure 9.4 Matches by productivity type in the Kagel and Roth (2000) experiment

Note. The figure shows the average cost of early matches (used as a metric, since the cost is higher the earlier the match) observed in each subgroup of five matching markets, separately for high- and low-productivity agents. After the first 10 matching markets, the third market period implements either the Newcastle or the DA matching algorithm.

Source: Kagel and Roth (2000, p. 217, Figure 2).

shown in Figure 9.4, this unravelling is mostly due to low-productivity types in the second period, who face greater incentives to unravel given that they are less attractive.

The introduction of a centralised clearing house reduces the number of early matches, whatever the matching algorithm. The adjustment dynamics are, however, very different between the two mechanisms. Figure 9.4 further disaggregates the data into

high- and low-productivity types. Early matches in periods 1 and 2 are pooled here using the cost of the early match – which is higher the earlier the match occurs. Under the Newcastle algorithm, the overall cost of early matches increases, while it is sharply reduced under the DA algorithm – the threat of facing the Newcastle algorithm thus increases unravelling, while DA reduces it. Again, this differential trend is mainly due to different kinds of behaviour according to productivity type. The cost of early matches steadily, and significantly, falls for high-productivity types under the DA algorithm, while it increases under the Newcastle algorithm. After five centralised experimental markets, high-productivity workers refuse all first-period early offers under the DA algorithm, while the acceptance rate remains at around 33% under the Newcastle algorithm.[4] These observations clearly show that the DA algorithm tends to eliminate mismatches, while the Newcastle algorithm rather encourages them. The results in the third period also show that the Gale–Shapley algorithm generates a sharp reduction in the number of unmatched agents (see Kagel and Roth, 2000, p. 219, Figure 3), and reduces the number of unstable matches, as compared to the Newcastle algorithm.

The experimental environment is thus a powerful tool for our understanding of two-sided matching markets. These results suggest that the DA algorithm will produce better allocations, by reducing unravelling and producing better matches. These results have been extended by further work in a number of directions: Nalbantian and Schotter (1995) consider the case of transferable utility between partners – the benefits from the match are thus no longer partner-specific but can be shared between them; Haruvy and Ünver (2007), Echenique and Yariv (2013) and Pais et al. (2012) focus on the stability of the matching arising in repeated decentralised markets; Chen and Sönmez (2006), Pais and Pintér (2008), Featherstone and Niederle (2016) and Echenique et al. (2016) widen the scope of matching algorithms (including, e.g. the Boston or the top trading cycle algorithm) and address a second major policy challenge in matching markets: whether agents have an incentive to manipulate, or truthfully reveal, their private information on the market.

9.4 Whispering in the Ear of Princes: Behavioural Public Policy

Decision-makers are interested in what would occur in a given situation were they to implement a different set of rules regarding how people interact and make decisions. They do not care about how universal/elegant/fruitful for further research is the framework in which behaviour is analysed. As a result, the variety of behaviours observed in experiments has quickly attracted the attention of decision-makers and become a central topic in the analysis of public policies – even if their exact theoretical implications largely remain an open question, as discussed in Section 9.1. The documented failures of the rationality assumptions render the theoretical debate about the founding principles of

[4] Moreover, all high-productivity firms make a second-period early offer after 10 rounds under the Newcastle algorithm.

public policies more complicated than it used to be. This point has been emphasised in the recent literature on 'liberal paternalism', giving rise to the new policy tools known as 'nudges'. This strand of literature is by far the best known of the recent advances in this area, and has provoked a great deal of debate. One crucial issue is that liberal paternalism takes departures from rational behaviour as granted, and carries out design choices based on the social planner's view of preferences. This leads us to think differently about the welfare analysis of public policies, which is the question that ends this section.

9.4.1 Liberal Paternalism: Liberal or Paternalistic?

There are two opposing traditions in the design of public policies.[5] On the one hand, paternalism compels individuals to make choices that are considered to be the best for them. The social planner decides on the final allocation of resources, and individuals are expected to comply with this centralised allocation. There are two standard arguments against this tradition in economics. First, the social planner has to be assumed to be benevolent; second, paternalism requires an enormous amount of information about individual preferences and the resources that are available to define the welfare function used to determine the final allocation. These criticisms are the main arguments in favour of liberalism, where freedom is best as it allows individuals to make choices in their own best interest.

The most important change over the past two decades has been that behavioural economics has shown that people are in fact very often not able to decide in their own best interest. As an example, under the status quo bias (see Chapter 5, Illustration 5.4) individuals who are statistically identical will make different decisions depending on which option is framed as the status quo. Since at most one option can be best for each decision-maker, decision-makers who are affected by the status quo bias are unable to decide what is best for them due to the framing of the situation.

This observation is the starting point for the recent discussions about 'liberal paternalism' (Thaler and Sunstein, 2003). As cognitive biases prevent individuals from taking the decisions that they intrinsically wish to take, public policies should be designed to help them in this respect. More precisely, liberal paternalism assumes that individuals know what is best for them (this is the liberal part of the approach) but that decision biases render them unable to make this choice. Public policies are thus expected to design incentives taking these biases into account, to allow individuals to take the decision they would have made had they been able to decide rationally. This second part is the paternalistic component, as public policies have to rely on some particular view of individuals' 'true' preferences and the choice that they would have made. The public-policy tools to take into account actual decision-making are

[5] The discussion in this section is far too short to give this topic the attention it deserves. We apologise for the many implicit assumptions we need to make to keep this presentation short and simple, and refer the reader to specialised contributions for a more appropriate coverage of the topic – see e.g. Li et al. (2014) for a literature review and insightful examples.

called 'nudges' (see Focus 9.3 for an overview), defined by Thaler and Sunstein (2008, p. 6) as:

'any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates. Putting fruit at eye level counts as a nudge. Banning junk food does not.'

Unsurprisingly, this approach has been criticised as sharing the drawbacks of both liberal (Munnell, 2003) and paternalistic (Mitchell, 2005) policies, leading many authors to consider it nothing more than an oxymoron. Sunstein and Thaler's (2003) response to these criticisms focuses on what they consider to be two misconceptions. The first is that paternalism is seen by liberals as avoidable. This would be true if institutions could be designed in which decisions are perfectly free. However, most choices involve unavoidable features that are known to influence choice. For instance, default rules (that apply to applications such as marriage, health care or retirement plans, as illustrated in Section 9.4.2 below) or starting values lead to anchoring as individuals overly rely on the first piece of information they receive. As such, the framing of the decision influences the decision itself. We thus have to make a choice regarding the architecture that is actually implemented. The typical example, first given by Sunstein and Thaler, is that of the placement of meals in a self-service restaurant. People are hungry when they enter and are thus more likely to choose the dishes that appear first in the line. Of course, there has to be an order, but there are different orders, e.g. that which maximises welfare, a random choice of order, or that which maximises obesity. If obesity reflects inconsistency and reduces individual well-being, liberal paternalism recommends that we should help them resist temptation by increasing transaction costs (for example, putting desserts at the end and vegetables first).

The second misconception, according to Sunstein and Thaler, is to consider paternalism as coercive. They rather argue that nudges induce an 'oriented free choice', and hence maintain the freedom associated with liberal policies. Both of these answers leave open the question of how the orientation can best be designed, i.e. (i) how actual mistakes can be disentangled from free choices (smoking might reflect time inconsistency, or just as well a perfectly rational decision according to some authors, e.g. Becker and Murphy, 1988) and (ii) how the social planner's preferences can be defined without knowing the true preference distribution in the population.[6] We illustrate the practical working of this approach in one of the most famous examples of a liberal paternalistic policy: the default option in retirement plans.

9.4.2 An Example: Time Inconsistency and the Default Option

As discussed in detail in Chapter 6, Section 6.6, efficient inter-temporal decisions require two central conditions on time preferences that are often found to be lacking in experiments: stationarity and dynamic consistency (Strotz, 1955). Stationarity implies that a decision made at date t does not change if all consequences of the choice are

[6] Sunstein and Thaler suggest two avenues: relying on cost benefit analysis, or adopting the solution that a majority would vote for – provided that votes are rational.

Focus 9.3
Designing a liberal and paternalistic choice architecture

Choice architecture, defined as the design of the context in which people make decisions (Thaler and Sunstein, 2008), is the central instrument of liberal paternalistic interventions. Johnson et al. (2012) group the tools available into two broad categories, as shown in the table below (from Table 1, p. 489).

Problem	Choice-architecture tools	Examples
	Amount of information available to the decision-maker	
Alternative overload	Reduce number of alternatives	Medicare drug plans (Kling et al., 2012), investments (Cronqvist and Thaler, 2004)
	Technology and decision aids	Sorting on attributes (Lynch Jr and Ariely, 2000); mobile devices and applications (Cook and Song, 2009); smart energy grids, decision inertia, use default investments (Cronqvist and Thaler, 2004; Madrian and Shea, 2001); insurance (Johnson et al., 1993); organ donation (Johnson and Goldstein, 2003)
Myopic procrastination	Focus on satisfying	Planning errors (Koehler, 1991; Weber and Lindemann, 2011; Shu, 2008), job search (Iyengar et al., 2006)
	Limited time windows	gift certificates (Shu and Gneezy, 2010), retirement planning (O'Donoghue and Rabin, 1998), tax credits
Long search process	Decision staging	Automobile customisation (Levav et al., 2010), product evaluation (Häubl et al., 2010)
	Framing of the decision	
Naive allocation	Partitioning of options	Investments Langer and Fox (2005); Bardolet et al. (2011), food menus, automobile attributes (Martin and Norton, 2009)
Attribute overload	Attribute parsimony and labelling	Good/bad labels for numeric information (Peters et al., 2009)
Non-linear attributes	Translate and rescale for better evaluability	Credit card repayments (Soll et al., 2013), fuel-mileage ratings (Larrick and Soll, 2008)

The first category, in the top panel of the table, refers to the amount of information provided to the decision-maker. For a given amount of information, different decisions can result from the way in which the decision is framed. This is the focus of the bottom part of the table. For each of the two categories, the first column describes the decision-making feature to be addressed, the second how the choice architecture can be adjusted to take the decision consequences into account, and the third a list of empirical applications in the literature.

delayed by the same amount of time. Under stationarity, the choice between two timed outcomes thus only depends on the time distance between them, resulting in constant impatience. Experimental evidence rather finds decreasing impatience (e.g. Frederick et al., 2002), or present bias: more weight is given to the immediate present than to similarly delayed future events. From a behavioural point of view, present-biased preferences reflect a lack of self-control: the decisions people would like to make in the future do not match their current behaviour.[7] Self-control problems can explain a variety of real-world behaviour, such as smoking, over-response to pre-teaser interest rates on credit cards, the overestimation of future health-club attendance and distorted views of deadlines (see DellaVigna, 2009, for a survey).

Saving is a typical inter-temporal high-stake decision that is faced by most individuals. Due to self-control problems, people might lack the ability to reduce present consumption in order to raise future consumption, for example during retirement. Combined with the difficulty of the planning task and bounded rationality, household saving rates may not be optimal.[8] A series of papers has proposed liberal paternalistic solutions to this kind of issue based on default rules.

Madrian and Shea (2001) consider the saving behaviour of employees at a large US firm when changing the enrolment rule in the national pension savings plan, the '401(k)'. Employees used to have to explicitly enrol, whereas now enrolment in the company's 401(k) plan for newly hired employees is automatic, unless they opt out. As discussed in Section 9.4.1, the status quo bias predicts that individuals will stick to the default option, even when there is no cost to changing to a better option. A change in the default is thus a typical tool in liberal paternalism, as people continue to have the choice to decide what suits them best, but are encouraged to make the decision that is best for them. One key point in preserving the libertarian nature of the 'default setting' is the cost of opting out: freedom of choice is conserved only when this is kept low (Thaler and Sunstein, 2008).

Madrian and Shea observed a dramatic rise in 401(k) participation by new workers 15 months after the introduction of this new default option, from 37% to 86% after the introduction of the automatic enrolment rule, leading to an 11% increase in overall 401(k) participation. Choi et al. (2004) have replicated the strong effect of automatic enrolment found by Madrian and Shea (2001) in three different companies. The results observed

[7] As a result, this kind of behaviour has also been explained by models of multiple selves (Benabou and Pycia, 2002; Bernheim and Rangel, 2004; Fudenberg and Levine, 2006). In the saving example, time consistency assumes that all selves agree on the saving plan. Under time inconsistency, on the contrary, the decision-maker can no longer trust himself to limit (or increase) consumption in the future (Eeckhoudt et al., 2005; Salanié and Treich, 2006): the self that consumes at time t has different preferences to the self that planned the saving beforehand.

[8] Although financial education has sometimes been suggested as a policy tool to circumvent the issues raised by bounded rationality in planning, it does not address self-control problems. Health insurance decisions are a typical example for which financial education plays a crucial role. For example, Bhargava et al. (2015) analyse the health-insurance decisions of employees at a large US firm where a new menu plan included a large share of financially dominated options. They find that a majority of employees make dominated choices. They underline the considerable negative consequences of these dominated choices, with excess spending of an average of 42% of the annual plan premium. In a series of hypothetical-choice experiments, the authors show that the lack of understanding of basic health insurance concepts plays a fundamental role in the observed choices.

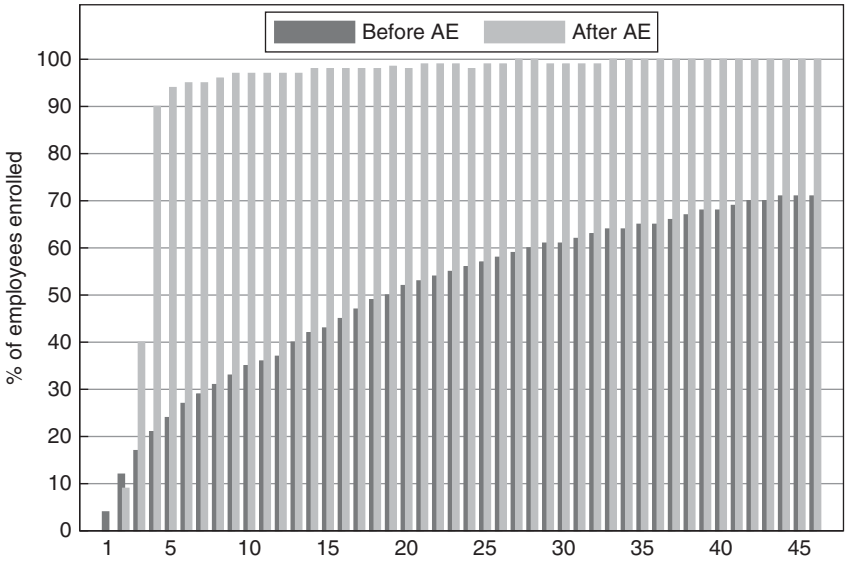


Figure 9.5 401(k) participation by tenure in Company A in Choi et al. (2004)

Note. For each tenure level, the figure shows the share of firm employees who enrolled in the 401(k) according to whether the default is to opt in or to opt out.
Source: Choi et al. (2004, p. 89, Figure 2.1).

in one of the three are shown in Figure 9.5, displaying the share of employees enrolled before and after the switch to automatic enrolment as a function of tenure in the firm. Before the introduction of automatic enrolment, the employee participation rate started at under 10% and increased progressively with tenure to less than 70%. Under automatic enrolment, the participation rate continued to start at a quite low level, but jumped after three to four months' tenure to 92%, and increased with tenure to almost 98%. These results more generally hold for all three of the firms analysed: 401(k) participation rates before automatic enrolment ranged from 26% to 69%, depending on the employees' tenure, and jumped to over 85% once participation was made the default choice.

Thaler and Benartzi (2004) consider another change in the default, specifically aimed at overcoming self-control problems. Their field experiment focuses on a programme called 'Save More Tomorrow', in which participants are offered the 401(k) plan with a very low initial contribution that increases over time – future pay raises are allocated to the saving plan. Most employees (78%) subscribe to the plan, which results in a rise from 3.5% to 13.6% in savings. These strong default-option effects are hard to understand, unless an unrealistically large psychological transaction cost is associated with switching. O'Donoghue and Rabin (2001, 1998) attempt to explain this effect. They find that the strong effect of the default option is due to self-control problems in the first place, which, combined with even a small amount of naivety, lead to potentially infinite procrastination.

Overall, this evidence is consistent with small and simple changes in the default having large consequences on behaviour. As such, behavioural policies might be much

more effective than traditional policies – typically based on capital income tax. For example, the US government spends around $100 billion per year in subsidising retirement saving in 401(k) and IRA accounts (Chetty, 2015). Chetty et al. (2014) estimate on Danish data that a one-dollar increase in defaults raises total savings by $0.85. As default effects have been observed over a wide range of decisions, from organ donation to car insurance plans and the consent to collect personal information, the paternalistic design of the default option may be broadly applicable. The effect of default options should not, however, be generalised too far, with a risk of overestimation. In further analysis of their data, Choi et al. (2004) show that, while automatic enrolment encourages 401(k) participation, it can anchor participants in conservative investment strategies and low saving rates: automatic enrolment has little positive impact on average long-run wealth accumulation. The main effect of automatic enrolment is to reduce the variance of wealth accumulation by drastically cutting the fraction of employees with no savings. In this case, the main effect of a change in the default is to increase the lower bound of the distribution of wealth accumulation. Moreover, setting a unique default is optimal only if a large number of employees share the same optimum, and if the default choice coincides with this optimum. With more heterogeneity in employees' preferences and personal characteristics, the welfare gains associated with a given default may be much lower. As described in Focus 9.4, an alternative is to force individuals to actually make decisions with a different choice situation architecture.

9.4.3 Welfare Evaluations

Rationality is not only a key requirement for free choices to match individually optimal outcomes, it also allows us to carry out welfare analysis based on the observed choices. In standard welfare analysis, with rational agents, the preferences revealed by choice are assumed to match the decision-maker's normative preferences perfectly. An economic agent who decides not to save is assumed to have made an informed choice, revealing their preference for current over future outcomes: not saving is the optimal outcome given these underlying individual preferences. Revealed preferences can thus be embedded in welfare analysis and used to compare outcomes and reach conclusions regarding the best decisions from the point of view of everyone's preferences – and hence what policies should be implemented.

This reasoning no longer holds under departures from rationality (McQuillin and Sugden, 2012). If the saving behaviour of this same decision-maker comes about due to time-inconsistent preferences, it is no longer the case that this outcome is strictly preferred from the agent's point of view. There is, however, ambiguity here: clearly, the current choice of a time-inconsistent decision-maker is their best current outcome – as otherwise they would have made another decision. But it is just as clearly not preferred for this same individual from the point of view of their future preferences over future outcomes. Beshears et al. (2008) describe five kinds of situation in which normative preferences are likely to diverge from the preferences revealed by decision-makers' choices.

Focus 9.4

Opt-in/opt-out versus active decisions: a non-liberal-paternalistic tool to enhance enrolment in 401(k) without default

In the two default options described in the text – automatic enrolment leading to an opt-out choice and non-participation leading to an opt-in choice – the fallback position defines the passive choice of the employee. Carroll et al. (2009) consider a third possibility, in which a formal choice of enrolment is required from the employee. There is no longer a default since employees must declare their enrolment preference in an active way. Paternalism is here limited to a cognitive and temporally limited intervention: employees are forced to think about the economic problem and to make a decision, with no answer to the problem being proposed by default. Carroll et al. (2009) collect data from one firm that switched unintentionally from an active-decision type to a standard type. The results show a drop in the enrolment rate from 69% under active decision to 41% under an opt-in choice. Conditional on demographics, the opt-in situation delayed the enrolment decision by 30 months, whereas active decisions forced the employees to decide immediately. The average saving rate and accumulated balances drastically increased thanks to this earlier enrolment. The authors also propose a theoretical framework to evaluate the welfare consequences of the different types of enrolment. The model shows that active decision is optimal when employees are present-biased, as active decisions force employees to counterbalance the effects of procrastination. Bernheim et al. (2015) show that defaults and penalties for passive choice are substitutable policy instruments. The optimal default policy is either an attractive default associated with no penalty, or an extremely high penalty with no matter which default.

1. *Passive choice*. This occurs when decision-makers face defaults that are set by some third party, and passively accept these defaults even if they do not correspond to their best option.
2. *Complexity*. Increased complexity is likely to generate noisier choices, complexity aversion and more delayed choices.
3. *Limited personal experience*. Decision-makers are more likely to learn from their own experience than from the experience of others. As a result, a lack of sufficient personal experience might limit the ability to learn, and so choose the best option.
4. *Third-party marketing* can also bias true preferences by manipulating information.
5. *Inter-temporal choice* is a standard choice environment in which decision-makers' choices are inconsistent.

When choice varies with arbitrary elements of the decision-making environment, the preferences revealed by observed behaviour will not be the same as normative preferences. In the words of Kahneman et al. (1997), 'experienced utility' – which corresponds to Bentham's concept of utility, also called 'true utility' by Bernheim et al. (2009) – is then different from 'decision utility' – the function that is maximised when decisions are made. Decision utility here, then, no longer reflects well-being, so that the use of revealed preferences and decision utility as normative tools is logically inconsistent (Goldin and Reck, 2015). There are, however, a variety of methods that allow us to recover the decision-makers' true preference based on their observed choices. For

example, Beshears et al. (2008) describe six different ways of identifying normative preferences under systematic decision-making errors.

1. *Structural estimation* of a behavioural model with a parameter vector θ and a set of normative axioms that map the parameter vector into normative preferences. While the behavioural model allows us to estimate all the components of the decision, the set of normative axioms restricts the parameter space to the normative parameters only, and excludes all of the non-normative estimated parameters. For example, Paserman (2008) estimates the degree of hyperbolic discounting in a job search model and derives the optimal policy interventions aimed at reducing unemployment by eliminating estimated present bias from the set of relevant parameters.
2. *Active-decisions* choice architecture, forcing decision-makers to explicitly choose an option rather than opting for a default. Carroll et al. (2009), described in Focus 9.4, is an example in the context of 401(k) plan participation.
3. *Asymptotic choice* that enables decision-makers to amass sufficient experience to make more informed and so better choices than inexperienced decision-makers – see Illustration 9.4 above for an example applied to preference reversal.
4. *Aggregate revealed preferences* can reveal normative preference if decision-making errors are symmetrically distributed around 0. In that case, aggregating preferences over individuals reveals the normative preferences of the representative agent.
5. *Self-reported preferences* are often considered less informative than revealed preferences, as they are a form of cheap talk. They nonetheless inform us about preferences, and possibly do a better job than revealed preferences when there are departures from rationality. Chetty (2015) argues that this approach, applied to subjective well-being, has the same strengths and weaknesses as contingent evaluation (as discussed in Section 2.4 of Chapter 2). Further criticisms regarding measurement, identification and the aggregation of self-reported subjective well-being are set out in Bernheim (2009).
6. *Informed preferences* come either from external experts or from decision-makers who have particularly good training and education regarding the choice to be made. Since the preferences revealed by these agents are arguably more similar to normative preferences, they should be overweighted in measurement. Chetty et al. (2009), for instance, rely on this method to consider the implementation of commodity taxes.

The above approaches amount to measuring well-behaved normative preferences based on (assumed) badly behaved choices. An alternative is to revise the foundations of welfare analysis in the light of behavioural economics. Bernheim (2009) discusses two competing views for this normative framework. The first takes the standard preference-based approach and focuses on the decision-makers' true objectives. The normal welfare-analysis framework is thus generalised to models that rationalise non-standard behaviour. The main difficulty here is the number of alternative rationalisations that can explain a given choice pattern. Bernheim (2009) argues that this plurality will conceptually render identification of welfare criteria impossible.

The second approach, introduced by Bernheim et al. (2009) and Salant and Rubinstein (2008), generalises conventional choice-based welfare analysis to situations in which agents make non-standard decisions. The welfare criterion (the 'unambiguous

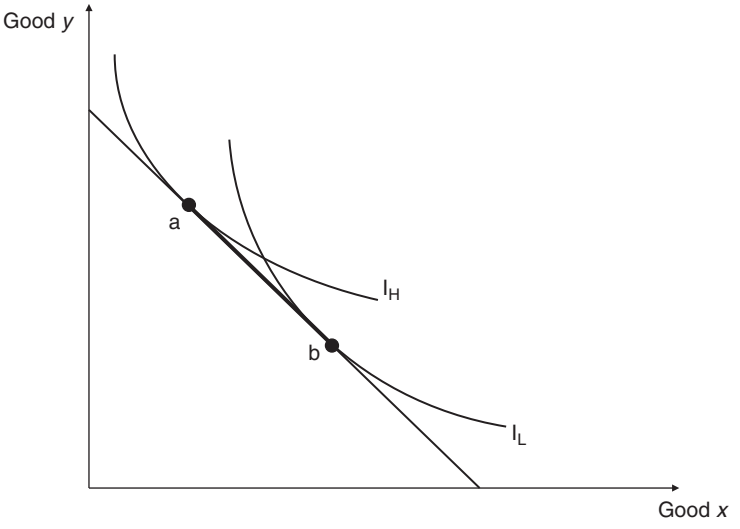


Figure 9.6 Individual welfare optima and consistent arbitrariness
Note. The figure shows the indifference curves and the budget constraint leading to optimal choices between good x and good y for two levels of anchoring, low (I_L) and high (I_H).
Source: Bernheim (2009, p. 303, Figure 1b).

choice relation') is defined so as to 'respect choice whenever it provides clear normative guidance, and to live with whatever ambiguity remains' (Bernheim et al., 2009, p. 53). Choices are assumed to reflect normative preferences whenever decision-makers make internally consistent decisions across the decision environments; the welfare relation is otherwise assumed to be incomplete. To produce normative guidance, one possibility is to construct welfare relations based on choice sets in which choices are consistent, and ignore the others. Bernheim et al. (2009) show that this welfare choice criterion is the only one to satisfy acyclicity (consistency), the respect of choice, and the fact that no option can be classified as a behavioural mistake based on choice patterns alone.

This incompleteness of the welfare function is a challenge for the usual tools used in welfare analysis, such as equivalent variations. When the standard axioms of choice hold, revealed preferences correspond to normative preferences and there is no ambiguity; when there are choice anomalies, by way of contrast, the welfare function only provides a range of equivalent variations. Figure 9.6 illustrates the welfare function associated with the consistent arbitrariness described in Focus 9.5. An individual s faces different anchors that influence choice over two goods, y and x. Two indifference curves are plotted along with the budget constraint: I_H is the indifference curve leading to optimal choice a under the high anchor, and I_L that leading to optimal choice b under the low anchor. Quantitative measures like equivalent variations will not be unique here, and are instead defined over a range of possible values. The set of individual welfare optima $[a, b]$ converges to a single utility-maximising choice as the anchoring effect is reduced. The standard welfare criterion thus appears as a limiting case of this more general setting, and can still be applied if the resulting ambiguity is small enough. It is only when large changes in behaviour are produced by the

Focus 9.5

The malleability of consumer preferences: anchoring and consistent arbitrariness

A number of experiments have shown that decision-makers can be very sensitive to salient but irrelevant features of the decision environment. Anchoring is a typical example, as highlighted by Tversky and Kahneman (1974). In this experiment, subjects first draw a random number (the 'anchor') and then face a judgement task involving a numerical assessment – for instance, guessing whether the number of African nations in the United Nations is greater or lower than the randomly drawn number. The results show that subjects' estimates are correlated with the arbitrary number drawn at the beginning of the experiment. Examples abound in the literature on the influence of anchors on judgements (see Epley and Gilovich, 2006, for references) but also on the elicitation of certainty equivalents for lotteries (Johnson and Schkade, 1989) or willingness to pay for public goods (Green et al., 1998). Ariely et al. (2003) go beyond subjective judgement and observe that a simple anchoring manipulation also affects preferences in an experiment with standard consumption goods (computer accessories, bottles of wine, luxury chocolate and books). In this experiment, subjects face a series of goods with almost the same retail price (around $70) and are first asked whether they would buy each good for a price that is equal to the last two digits of their social security number. After this anchoring task, subjects are asked to state their willingness to pay for each good. At the end of the experiment, one task is selected at random and played for real to determine payoffs based on a Becker–DeGroot–Marshak mechanism. The following table (from Ariely et al., 2003, p. 76, Table 1) lists the average willingness to pay for each good, by the five quintiles of the social-security number distribution. The last column shows the correlation between these numbers.

Quintile	1	2	3	4	5	Correlation
Cordless trackball	8.64	11.82	13.45	21.18	26.18	0.415
Cordless keyboard	16.09	26.82	29.27	34.55	55.64	0.516
Average wine	8.64	14.45	12.55	15.45	27.91	0.328
Vintage wine	11.73	22.45	18.09	24.55	37.55	0.319
Design book	12.82	16.18	15.82	19.27	30.00	0.419

Whatever the good considered, willingness to pay is significantly correlated with the social-security number used as an anchor. The valuations in the top quintile are three times larger than those in the bottom quintile. The absolute valuations of the goods thus appear to a large extent arbitrary. The large majority of subjects, however, exhibit stable relative preferences: the value of the cordless keyboard is twice as high as the cordless trackball, and vintage wine is valued more than average wine. The relative valuations are thus consistent with demand curves derived from fundamental preferences. The experiment then suggests that valuations combine both arbitrariness and consistency. Bergman et al. (2010) replicate these findings on willingness to pay, and Fudenberg et al. (2012) find small effects on both willingness to pay and willingness to accept.

anchor that the unambiguous choice relation used to evaluate welfare will differ notably from a standard choice relation. In this case, the range of possible equivalent variations is wide, and the welfare criterion no longer allows any fine distinctions between the different allocations.

Bernheim et al. (2015) apply behavioural welfare analysis to study the welfare effects of default options in the 401(k) case. They consider a number of theories of default effects and evaluate their welfare consequences through two equivalent-variation measures associated with the change in policy. The first is the smallest increment in income, in the context of the initial policy, such that the chosen bundle of goods under the initial policy will unambiguously be chosen over the bundle obtained with the new policy. The second is the largest increment in income, in the context of the new policy, such that the chosen bundle of goods under the new policy is unambiguously chosen over the bundle obtained with the initial policy. These two measures provide bounds on the equivalent variations, as the policy change is worth no more than the first measure and at least as much as the second. The analysis shows that the design of optimal welfare policy depends critically on the underlying behavioural model of choice. When time consistency or inattention matters, there is only little ambiguity about the normative effects of default rules, even if the opt-out costs that rationalise standard behaviour are fairly large in size. On the contrary, the degree of normative ambiguity for models with anchoring is much larger. For models of time consistency, the results show that minimising the opt-out cost by setting the default equal to the employer matching contribution cap is a suitable welfare-maximising policy – in line with the results of Sunstein and Thaler (2003). Last, the optimal default rate for models of anchoring is zero.

Summary

This chapter has focused on two questions. The first asked what should be done with theories that have been widely and convincingly tested in the lab. The second is related to the predictive value of experimental results from a public-policy perspective. The first question stems from the observation of thirty years of investigations in experimental economics. Accumulated evidence has renewed and broadened the traditional view of the determinants of economic behaviour. This accumulation of results has helped economists to determine which elements of their toolbox were accurate and which were not. In Section 9.1, we described how experimental economics has renewed the debate over the status of models and theory in economics. In particular, we showed how experimental results can help the economist to build better models, with an improved understanding of human behaviour. The evidence accumulated has also raised new questions about the necessity (or lack of necessity) to improve models. Through various examples and illustrations, this chapter has discussed from which kind of data it is legitimate to carry out induction and thereby improve models. We show that behavioural failures might not always imply theoretical failure, but rather a failure in the range of situations to which the theory applies. In Section 9.2, a case study on level-k reasoning illustrated how these questions apply to game theory.

The second part of the chapter focused on the public-policy consequences of the decision-making processes observed in the lab. The accumulated experimental results on rational behaviour show a balanced picture. Depending on the choice situations, institutions and incentives, the hypothesis of rational agents making decisions maximising their utility is confirmed. However, in many instances, agents are unable to conform or commit to these decisions. These situations have been systematically and extensively studied by what is often called behavioural economics. Section 9.3 takes the example of matching markets to show how the methodology of experimental economics can be used to enhance our understanding of markets and regulations, using simple decision-making in the lab. Section 9.4 discusses the public-policy consequences of the (experimentally) documented failures of rationality assumptions. The recognition of these failures has led to a complete renewal of thinking in public policy. This renewal has called the liberal view, in which individuals are able to choose what is best for them, into question. Section 9.4 presented and discussed the central element of this renewal, called liberal paternalism. Section 9.4 also presented some of the consequences of the violations of rationality in terms of traditional welfare analysis.

References

Aadland, David, and Arthur J. Caplan (2006) 'Cheap talk reconsidered: new evidence from CVM.' *Journal of Economic Behavior & Organization* 60(4), 562–78

Aadland, David, Arthur J. Caplan and Owen Phillips (2007) 'A Bayesian examination of information and uncertainty in contingent valuation.' *Journal of Risk and Uncertainty* 35(2), 149–78

Abbink, Klaus, and Heike Hennig-Schmidt (2006) 'Neutral versus loaded instructions in a bribery experiment.' *Experimental Economics* 9(2), 103–21

Abdellaoui, Mohammed (2000) 'Parameter-free elicitation of utility and probability weighting functions.' *Management Science* 46(11), 1497–1512

Abdellaoui, Mohammed, Arthur E. Attema and Han Bleichrodt (2010) 'Intertemporal tradeoffs for gains and losses: an experimental measurement of discounted utility.' *Economic Journal* 120(545), 845–66

Abdellaoui, Mohammed, Aurélien Baillon, Laetitia Placido and Peter P. Wakker (2011) 'The rich domain of uncertainty: source functions and their experimental implementation.' *American Economic Review* 101(2), 695–723

Abdellaoui, Mohammed, Carolina Barrios and Peter P. Wakker (2007a) 'Reconciling introspective utility with revealed preference: experimental arguments based on prospect theory.' *Journal of Econometrics* 138(1), 356–78

Abdellaoui, Mohammed, Han Bleichrodt, Emmanuel Kemel and Olivier L'Haridon (2016a) 'Beliefs and attitudes for natural sources of uncertainty.' *Working paper*

Abdellaoui, Mohammed, Han Bleichrodt and Olivier L'Haridon (2008) 'A tractable method to measure utility and loss aversion under prospect theory.' *Journal of Risk and Uncertainty* 36(3), 245–66

Abdellaoui, Mohammed, Han Bleichrodt, Olivier L'Haridon and Dennie van Dolder (2016b) 'Measuring loss aversion under ambiguity: a method to make prospect theory completely observable.' *Journal of Risk and Uncertainty* 52(1), 1–20

Abdellaoui, Mohammed, Han Bleichrodt, Olivier L'Haridon and Corina Paraschiv (2013) 'Is there one unifying concept of utility? An experimental comparison of utility under risk and utility over time.' *Management Science* 59(9), 2153–69

Abdellaoui, Mohammed, Han Bleichrodt and Corina Paraschiv (2007b) 'Loss aversion under prospect theory: a parameter-free measurement.' *Management Science* 53(10), 1659–74

Abdellaoui, Mohammed, Franck Vossmann and Martin Weber (2005) 'Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty.' *Management Science* 51(9), 384–99

Agranov, Marina, Elizabeth Potamites, Andrew Schotter and Chloe Tergiman (2012) 'Beliefs and endogenous cognitive levels: an experimental study.' *Games and Economic Behavior* 75(2), 449–63

Ahlbrecht, Martin, and Martin Weber (1997) 'An empirical study on intertemporal decision making under risk.' *Management Science* 43(6), 813–26

Ajzen, Icek, Thomas C. Brown and Franklin Carvajal (2004) 'Explaining the discrepancy between intentions and actions: the case of hypothetical bias in contingent valuation.' *Personality and Social Psychology Bulletin* 30(9), 1108–20

Akerlof, George A. (1982) 'Labor contracts as partial gift exchange.' *Quarterly Journal of Economics* 97(4), 543–69

Akerlof, George A., and Janet L. Yellen (1990) 'The fair wage-effort hypothesis and unemployment.' *Quarterly Journal of Economics* 105(2), 255–83

Akiyama, Eizo, Nobuyuki Hanaki and Ryuichiro Ishikawa (2017) 'It is not just confusion! Strategic uncertainty in an experimental asset market.' *Economic Journal*, forthcoming

Alekseev, Aleksandr, Gary Charness and Uri Gneezy (2017) 'Experimental methods: When and why contextual instructions are important.' *Journal of Economic Behavior & Organization* 134, 48–59

Allais, Maurice (1953) 'Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école americaine.' *Econometrica* 21(4), 503–46

Al-Ubaydli, Omar, and John A. List (2015) 'Do natural field experiments afford researchers more or less control than laboratory experiments?' *American Economic Review* 105(5), 462–66

American Psychology Association, APA (2008) 'Reporting standards for research in psychology: why do we need them? What might they be?' *American Psychologist* 63(9), 839–51

Ami, Dominique, Frederic Aprahamian, Olivier Chanel and Stephane Luchini (2011) 'A test of cheap talk in different hypothetical contexts: the case of air pollution.' *Environmental & Resource Economics* 50(1), 111–30

Amir, Ofra, David G. Rand and Ya'akov Kobi Gal (2012) 'Economic games on the internet: the effect of $1 stakes.' *PLoS ONE* 7(2), e31461

Anderhub, Vital, Werner Güth, Uri Gneezy and Doron Sonsino (2001b) 'On the interaction of risk and time preferences: an experimental study.' *German Economic Review* 2(3), 239–53

Anderhub, Vital, Rudolf Muller and Carsten Schmidt (2001a) 'Design and evaluation of an economic experiment via the internet.' *Journal of Economic Behavior & Organization* 46(2), 227–47

Andersen, Steffen, Seda Ertac, Uri Gneezy, Moshe Hoffman and John A. List (2011) 'Stakes matter in ultimatum games.' *American Economic Review* 101(7), 3427–39

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau and Elisabet E. Rutström (2008) 'Eliciting risk and time preferences.' *Econometrica* 76(3), 583–618

—— (2009) 'Elicitation using multiple price list formats.' *Experimental Economics* 12(3), 365–6

Anderson, Jon, Stephen V. Burks, Jeffrey Carpenter, Lorenz Goette, Karsten Maurer, Daniele Nosenzo, Ruth Potter, Kim Rocha and Aldo Rustichini (2013) 'Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples.' *Experimental Economics* 16(2), 170–89

Anderson, Lisa R., and Jennifer M. Mellor (2009) 'Are risk preferences stable? Comparing an experimental measure with a validated survey-based measure.' *Journal of Risk and Uncertainty* 39(2), 137–60

Andreoni, James (1995) 'Cooperation in public-goods experiments: kindness or confusion?' *American Economic Review* 85(4), 891–904

Andreoni, James, Michael A. Kuhn and Charles Sprenger (2013) 'On measuring time preferences.' *NBER WP*

Andreoni, James, and Charles Sprenger (2012) 'Estimating time preferences from convex budgets.' *American Economic Review* 102(7), 3333–56

Angner, Eric (2012) *A Course in Behavioral Economics* (New York: Palgrave Macmillan)

Angrist, Joshua D., and Guido W. Imbens (1999) 'Comment on James J. Heckman, "Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations".' *Journal of Human Resources* 34(4), 823–27

Angrist, Joshua D., and Alan B. Krueger (1999) 'Empirical strategies in labor economics.' In *Handbook of Labor Economics,* ed. Orley C. Ashenfelter and David Card, vol. 3, Part A (Amsterdam: Elsevier), pp. 1277–1366

Anscombe, Francis J. (1973) 'Graphs in statistical analysis.' *American Statistician* 27(1), 17–21

Arad, Ayala, and Ariel Rubinstein (2012) 'The 11–20 money request game: a level-*k* reasoning study.' *American Economic Review* 102(7), 3561–73

Ariely, Dan, George Loewenstein and Drazen Prelec (2003) 'Coherent arbitrariness: stable demand curves without stable preferences.' *Quarterly Journal of Economics* 118(1), 73–105

Ariely, Dan, and Michael I. Norton (2007) 'Psychology and experimental economics: A gap in abstraction.' *Current Directions in Psychological Science* 16(6), 336–9

Armantier, Olivier, and Amadou Boly (2013) 'On the effects of incentive framing on bribery: evidence from an experiment in Burkina Faso.' *Economics of Governance* 15(1), 1–15

Armantier, Olivier, and Nicolas Treich (2009) 'Subjective probabilities in games: an application to the overbidding puzzle.' *International Economic Review* 50(4), 1079–1102

—— (2013) 'Eliciting beliefs: proper scoring rules, incentives, stakes and hedging.' *European Economic Review* 62, 17–40

Aron, Arthur, Elaine N. Aron and Danny Smollan (1992) 'Inclusion of other in the self scale and the structure of interpersonal closeness.' *Journal of Personality and Social Psychology* 63(4), 596–612

Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers and Eric Zitzewitz (2008) 'The promise of prediction markets.' *Science* 320(5878), 877–8

Arrow, Kenneth J., Robert Solow, Paul Portney, Edward E. Leamer, Roy Radner and Howard Schuman (1993) 'Report of the NOAA panel on contingent valuation.' *Federal Register* 58(10), 4601–14

Attanasi, Giuseppe, Nikolaos Georgantzís, Valentina Rotondi and Daria Vigani (2017) 'Lottery- and survey-based risk attitudes linked through a multichoice elicitation task.' *Theory and Decision*, forthcoming

Attema, Arthur E., Han Bleichrodt, Yu Gao, Zhenxing Huang and Peter P. Wakker (2016) 'Measuring discounting without measuring utility.' *American Economic Review* 106(6), 1476–94

Attema, Arthur E., Han Bleichrodt, Kirsten I. M. Rohde and Peter P. Wakker (2010) 'Time-tradeoff sequences for analyzing discounting and time inconsistency.' *Management Science* 56(11), 2015–30

Augenblick, Ned, Muriel Niederle and Charles Sprenger (2015) 'Working over time: Dynamic inconsistency in real effort tasks.' *Quarterly Journal of Economics* 130(3), 1067–1115

Aumann, Robert J. (1995) 'Backward induction and common knowledge of rationality.' *Games and Economic Behavior* 8(1), 6–19

—— (1998) 'On the centipede game.' *Games and Economic Behavior* 23(1), 97–105

Bach, Dominik R. (2016) 'Skin conductance measures in neuroeconomic research.' In *Neuroeconomics*, ed. Martin Reuter and Christian Montag (Berlin and Heidelberg: Springer), 345–57

Baillon, Aurélien (2008) 'Eliciting subjective probabilities through exchangeable events: an advantage and a limitation.' *Decision Analysis* 5(2), 76–87

Baillon, Aurélien, and Han Bleichrodt (2015) 'Testing ambiguity models through the measurement of probabilities for gains and losses.' *American Economic Journal: Microeconomics* 7(2), 77–100

Baillon, Aurélien, Han Bleichrodt, Umut Keskin, Olivier L'Haridon and Chen Li (2017) 'Learning under ambiguity: an experiment using initial public offerings on a stock market.' *Management Science*, forthcoming

Ball, Sheryl Beth, and Paula-Ann Cech (1996) 'Subject pool choice and treatment effects in economic laboratory research.' *Research in Experimental Economics: A Research Annual* 6, 239–92

Baltussen, Guido, G. Thierry Post, Martijn J. van den Assem and Peter P. Wakker (2012) 'Random incentive systems in a dynamic choice experiment.' *Experimental Economics* 15(3), 418–43

Baltussen, Guido, Martijn J. van den Assem and Dennie van Dolder (2016) 'Risky choice in the limelight.' *Review of Economics and Statistics* 98(2), 318–32

Barberis, Nicholas C. (2013) 'Thirty years of prospect theory in economics: a review and assessment.' *Journal of Economic Perspectives* 27(1), 173–96

Bardolet, David, Craig R. Fox and Daniel Lovallo (2011) 'Corporate capital allocation: a behavioral perspective.' *Strategic Management Journal* 32(13), 1465–83

Bardsley, Nicholas, Robin Cubitt, Graham Loomes, Peter G. Moffatt, Chris Starmer and Robert Sugden (2009) *Experimental Economics: Rethinking the Rules* (Princeton, NJ: Princeton University Press)

Barsky, Robert B., Miles S. Kimball, F. Thomas Juster and Matthew D. Shapiro (1997) 'Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement survey.' *Quarterly Journal of Economics* 112(2), 537–79

Battalio, Raymond C., Larry Samuelson and John B. van Huyck (2001) 'Optimization incentives and coordination failure in laboratory stag hunt games.' *Econometrica* 69(3), 749–64

Beattie, Jane, and Graham Loomes (1997) 'The impact of incentives upon risky choice experiments.' *Journal of Risk and Uncertainty* 14(2), 155–68

Beauchamp, Jonathan P., Daniel J. Benjamin, Christopher F. Chabris and David I. Laibson (2012) 'How malleable are risk preferences and loss aversion.' *Working paper*

Becker, Gary S., and Kevin M. Murphy (1988) 'A theory of rational addiction.' *Journal of Political Economy* 96(4), 675–700

Becker, Selwyn W., and Fred O. Brownson (1964) 'What price ambiguity? Or the role of ambiguity in decision-making.' *Journal of Political Economy* 72(1), 62–73

Beggs, Jodi N. (2013) 'Homer economicus or homer sapiens? Behavioral economics in The Simpsons.' *Working paper*

Bellemare, Charles, Luc Bissonnette and Sabine Kröger (2014) 'Statistical power of within and between-subjects designs in economic experiments.' *CESifo working paper*

Bellemare, Charles, and Sabine Kröger (2007) 'On representative social capital.' *European Economic Review* 51(1), 183–202

Bellemare, Charles, Sabine Kröger and Arthur van Soest (2008) 'Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities.' *Econometrica* 76(4), 815–39

Bellemare, Charles, and Bruce Shearer (2009) 'Gift giving and worker productivity: evidence from a firm-level experiment.' *Games and Economic Behavior* 67(1), 233–44.

Benabou, Roland, and Marek Pycia (2002) 'Dynamic inconsistency and self-control: a planner–doer interpretation.' *Economics Letters* 77(3), 419–24

Benabou, Roland, and Jean Tirole (2003) 'Intrinsic and extrinsic motivation.' *Review of Economic Studies* 70(3), 489–520

Benhabib, Jess, Alberto Bisin and Andrew Schotter (2010) 'Present-bias, quasi-hyperbolic discounting, and fixed costs.' *Games and Economic Behavior* 69(2), 205–23

Ben-Porath, Elchanan (1997) 'Rationality, Nash equilibrium and backwards induction in perfect-information games.' *Review of Economic Studies* 64(1), 23–46

Benzion, Uri, Amnon Rapoport and Joseph Yagil (1989) 'Discount rates inferred from decisions: an experimental study.' *Management Science* 35(3), 270–84

Berg, Joyce, John Dickhaut and Kevin McCabe (1995) 'Trust, reciprocity, and social history.' *Games and Economic Behavior* 10(1), 122–42

Bergman, Oscar, Tore Ellingsen, Magnus Johannesson and Cicek Svensson (2010) 'Anchoring and cognitive ability.' *Economics Letters* 107(1), 66–8

Bernheim, B. Douglas (2009) 'Behavioral welfare economics.' *Journal of the European Economic Association* 7(2–3), 267–319

Bernheim, B. Douglas, Andrey Fradkin and Igor Popov (2015) 'The welfare economics of default options in 401(k) plans.' *American Economic Review* 105(9), 2798–2837

Bernheim, B. Douglas, and Antonio Rangel (2004) 'Addiction and cue-triggered decision processes.' *American Economic Review* 94(5), 1558–90

—— (2009) 'Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics.' *Quarterly Journal of Economics* 124(1), 51–104

Beshears, John, James J. Choi, David I. Laibson and Brigitte C. Madrian (2008) 'How are preferences revealed?' *Journal of Public Economics* 92(8), 1787–94

Bhargava, Saurabh, George Loewenstein and Justin Sydnor (2015) 'Do individuals make sensible health insurance decisions? Evidence from a menu with dominated options.' *NBER working paper*

Binmore, Ken (1999) 'Why experiment in economics?' *Economic Journal* 109(453), 16–24

Binmore, Ken, and Avner Shaked (2010) 'Experimental economics: where next?' *Journal of Economic Behavior & Organization* 73(1), 87–100

Binmore, Ken, Lisa Stewart and Alex Voorhoeve (2012) 'How much ambiguity aversion?' *Journal of Risk and Uncertainty* 45(3), 215–38

Binswanger, Hans P. (1980) 'Attitudes toward risk: experimental measurement in rural India.' *American Journal of Agricultural Economics* 62(3), 395–407

Bjornstad, David, Ronald G. Cummings and Laura Osborne (1997) 'A learning design for reducing hypothetical bias in the contingent valuation method.' *Environmental & Resource Economics* 10(3), 207–21

Black, Sandra E. (1999) 'Do better schools matter? Parental valuation of elementary education.' *Quarterly Journal of Economics* 114(2), 577–99

Blamey, Russel K., Jeff W. Bennett and Mark Daniel Morrison (1999) 'Yea-saying in contingent valuation surveys.' *Land Economics* 75(1), 126–41

Blanco, Mariana, Dirk Engelmann, Alexander Koch and Hans-Theo Normann (2010) 'Belief elicitation in experiments: is there a hedging problem?' *Experimental Economics* 13(4), 412–38

Blanco, Mariana, Dirk Engelmann and Hans-Theo Normann (2011) 'A within-subject analysis of other-regarding preferences.' *Games and Economic Behavior* 72(2), 321–38

Bleichrodt, Han (2001) 'Probability weighting in choice under risk: an empirical test.' *Journal of Risk and Uncertainty* 23(2), 185–98

Bleichrodt, Han, Alessandra Cillo and Enrico Diecidue (2010) 'A quantitative measurement of regret theory.' *Management Science* 56(1), 161–75

Bleichrodt, Han, Yu Gao and Kirsten I.M. Rohde (2016) 'A measurement of decreasing impatience for health and money.' *Journal of Risk and Uncertainty* 52(3), 213–31

Bleichrodt, Han, and Jose Luis Pinto (2000) 'A parameter-free elicitation of the probability weighting function in medical decision analysis.' *Management Science* 46(11), 1485–96

Bleichrodt, Han, Jose Luis Pinto and Peter P. Wakker (2001) 'Making descriptive use of prospect theory to improve the prescriptive use of expected utility.' *Management Science* 47(11), 1498–1514

Bleichrodt, Han, Kirsten I.M. Rohde and Peter P. Wakker (2009) 'Non-hyperbolic time inconsistency.' *Games and Economic Behavior* 66(1), 27–38

Blundell, Richard (2010) 'Comments on: Michael P. Keane "structural vs. atheoretic approaches to econometrics".' *Journal of Econometrics* 156(1), 25–6

Bohm, Peter (1972) 'Estimating demand for public goods: an experiment.' *European Economic Review* 3(2), 111–30

Bolton, Gary E., and Axel Ockenfels (2000) 'Erc: a theory of equity, reciprocity, and competition.' *American Economic Review* 90(1), 166–93

Bonetti, Shane (1998) 'Experimental economics and deception.' *Journal of Economic Psychology* 19(3), 377–95

Booij, Adam S., and Gijs van de Kuilen (2009) 'A parameter-free analysis of the utility of money for the general population under prospect theory.' *Journal of Economic Psychology* 30(4), 651–66

Booij, Adam S., Bernard M.S. van Praag and Gijs van de Kuilen (2010) 'A parametric analysis of prospect theory's functionals for the general population.' *Theory and Decision* 68(1-2), 115–48

Borghans, Lex, Angela Lee Duckworth, James J. Heckman and Bas Weel (2008) 'The economics and psychology of personality traits.' *Journal of Human Resources* 43(4), 972–1059

Bosch-Domenech, Antoni, Jose G. Montalvo, Rosemarie Nagel and Albert Satorra (2002) 'One, two,(three), infinity, . . . : newspaper and lab beauty-contest experiments.' *American Economic Review* 92(5), 1687–1701

Bosch-Domènech, Antoni, and Joaquim Silvestre (2013) 'Measuring risk aversion with lists: a new bias.' *Theory and Decision* 75(4), 465–96

Bostic, Raphael, Richard J. Herrnstein and Duncan R. Luce (1990) 'The effect on the preference-reversal phenomenon of using choice indifferences.' *Journal of Economic Behavior & Organization* 13(2), 193–212

Boulu-Reshef, Béatrice, Irene Comeig, Robert Donze and Gregory D. Weiss (2016) 'Risk aversion in prediction markets: a framed-field experiment.' *Journal of Business Research* 69(11), 5071–75

Brandts, Jordi, and Gary Charness (2000) 'Hot vs. cold: sequential responses and preference stability in experimental games.' *Experimental Economics* 2(3), 227–38

—— (2011) 'The strategy versus the direct-response method: a first survey of experimental comparisons.' *Experimental Economics* 14(3), 375–98

Bremzen, Andrei, Elena Khokhlovaz, Anton Suvorov, and Jeroen van de Ven (2013) 'Bad news: an experimental study on the informational effects of rewards.' *Review of Economics and Statistics*, forthcoming

Brewer, Marilynn B., and William D. Crano (2014) 'Research design and issues of validity.' In *Handbook of Research Methods in Social and Personality Psychology,* ed. Harry T. Reis and Charles M. Judd (Cambridge: Cambridge University Press), 11–26

Brier, Glenn W. (1950) 'Verification of forecasts expressed in terms of probability.' *Monthly Weather Review* 78(1), 1–3

Brislin, Richard W. (1970) 'Back-translation for cross-cultural research.' *Journal of Cross-cultural Psychology* 1(3), 185–216

Brockwell, Sarah E., and Ian R. Gordon (2001) 'A comparison of statistical methods for meta-analysis.' *Statistics in Medicine* 20(6), 825–40

Brown, Thomas C., Icek Ajzen and Daniel Hrubes (2003) 'Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation.' *Journal of Environmental Economics and Management* 46(2), 353–61

Bruggen, Alexander, and Martin Strobel (2007) 'Real effort versus chosen effort in experiments.' *Economics Letters* 96(2), 232–6

Bruhin, Adrian, Helga Fehr-Duda and Thomas Epper (2010) 'Risk and rationality: uncovering heterogeneity in probability distortion.' *Econometrica* 78(4), 1375–1412

Bruni, Luigino, and Robert Sugden (2007) 'The road not taken: how psychology was removed from economics, and how it might be brought back.' *Economic Journal* 117(516), 146–73

Brunswick, E. (1956) *Perception and the Representative Design of Psychological Experiment* (Berkeley: University of California Press)

Buhrmester, Michael, Tracy Kwang and Samuel D. Gosling (2011) 'Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data?' *Perspectives on Psychological Science* 6(1), 3–5

Bull, Clive, Andrew Schotter and Keith Weigelt (1987) 'Tournaments and piece rates: an experimental study.' *Journal of Political Economy* 95(1), 1–33

Burchardi, Konrad B., and Stefan P. Penczynski (2014) 'Out of your mind: eliciting individual reasoning in one shot games.' *Games and Economic Behavior* 84, 39–57

Burks, Stephen V., Jeffrey Carpenter, Lorenz Goette and Aldo Rustichini (2012) 'Which measures of time preference best predict outcomes: evidence from a large-scale field experiment.' *Journal of Economic Behavior & Organization* 84(1), 308–20

Cadsby, C. Bram, Elizabeth Maynes and Viswanath Umashanker Trivedi (2006) 'Tax compliance and obedience to authority at home and in the lab: a new experimental approach.' *Experimental Economics* 9(4), 343–59

Camerer, Colin F. (1996) 'Rules for experimenting in psychology and economics, and why they differ.' In *Understanding Strategic Interaction: Essays in Honor of Reinhard Selten* (Berlin: Springer-Verlag), 313–27

—— (1998) 'Can asset markets be manipulated? A field experiment with racetrack betting.' *Journal of Political Economy* 106(3), 457–82

—— (2003) *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton, NJ: Princeton University Press)

—— (2015) 'The promise and success of lab–field generalizability in experimental economics: a critical reply to Levitt and List.' In *Handbook of Experimental Economic Methodology,* ed. Guillaume R. Fréchette and Andrew Schotter (New York: Oxford University Press), 249–95

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen and

Hang Wu (2016) 'Evaluating replicability of laboratory experiments in economics.' *Science* 351(6280), 1433–6

Camerer, Colin. F., Teck-Hua Ho and Juin-Kuan Chong (2004) 'A cognitive hierarchy model of behavior in games.' *Quarterly Journal of Economics* 119(3), 861–98

Camerer, Colin F., and Robin M. Hogarth (1999) 'The effects of financial incentives in experiments: a review and capital-labor-production framework.' *Journal of Risk and Uncertainty* 19(1), 7–42

Camerer, Colin F., Samuel Issacharoff, George Loewenstein, Ted O'Donoghue and Matthew Rabin (2003) 'Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism".' *University of Pennsylvania Law Review* 151, 1211–54

Camerer, Colin F., and George Loewenstein (2003) 'Behavioral economics: Past, present, future.' In *Advances in Behavioral Economics,* ed. Colin F. Camerer, George Loewenstein and Matthew Rabin (Princeton, NJ: Princeton University Press)

Campbell, Donald T. (1969) 'Reforms as experiments.' *American Psychologist* 24(4), 409–29

Campbell, Donald T., and Julian Stanley (1963) *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally)

Capen, Edward C., Robert V. Clapp and William M. Campbell (1971) 'Competitive bidding in high-risk situations.' *Journal of Petroleum Technology* 23, 641–53

Card, David, and Alan B. Krueger (1994) 'Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania.' *American Economic Review* 84(4), 772–93

Cardenas, Juan Camilo, Nicolas De Roux, Christian R. Jaramillo and Luis Roberto Martinez (2014) 'Is it my money or not? An experiment on risk aversion and the house-money effect.' *Experimental Economics* 17(1), 47–60

Cardenas, Juan Camilo, and Jeffrey Carpenter (2013) 'Risk attitudes and economic well-being in Latin America.' *Journal of Development Economics* 103, 52–61

Carlsson, Fredrik, and Peter Martinsson (2006) 'Do experience and cheap talk influence willingness to pay in an open-ended contingent valuation survey?' *Göteborg University, Department of Economics working paper*

Carpenter, Jeffrey, Allison Liati and Brian Vickery (2010) 'They come to play.' *Rationality and Society* 22(1), 83–102

Carroll, Gabriel D., James J. Choi, David I. Laibson, Brigitte C. Madrian and Andrew Metrick (2009) 'Optimal defaults and active decisions.' *Quarterly Journal of Economics* 124(4), 1639–74

Carson, Richard T., Theodore Groves, John A. List and Mark J. Machina (2002) 'Probabilistic influence and supplemental benefits: a field test of the two key assumptions underlying stated preferences.' *Department of Economics, University of California at San Diego working paper*

Carson, Richard T., Theodore Groves and Mark J. Machina (2000) 'Incentive and informational properties of preference questions.' *University of California at San Diego working paper*

Cartwright, Edward (2011) *Behavioral Economics* (New York: Routledge)

Chabris, Christopher F., David I. Laibson, Carrie L. Morris, Jonathon P. Schuldt and Dmitry Taubinsky (2008) 'Individual laboratory-measured discount rates predict field behavior.' *Journal of Risk and Uncertainty* 37(2-3), 237–69

Chamberlin, Edward H. (1948) 'An experimental imperfect market.' *Journal of Political Economy* 56(2), 95–108

Champ, Patricia A., Richard C. Bishop, Thomas C. Brown and Daniel W. McCollum (1997) 'Using donation mechanisms to value nonuse benefits from public goods.' *Journal of Environmental Economics and Management* 33(2), 151–62

Charness, Gary, Ernan Haruvy and Doron Sonsino (2007) 'Social distance and reciprocity: an internet experiment.' *Journal of Economic Behavior & Organization* 63(1), 88–103

Charness, Gary, and Peter Kuhn (2010) 'Lab labor: what can labor economists learn from the lab?' In *Handbook of Labor Economics,* ed. Orley Ashenfelter and David Card, vol. 4 (Amsterdam: Elsevier)

Charness, Gary, and Matthew Rabin (2002) 'Understanding social preferences with simple tests.' *Quarterly Journal of Economics* 117(3), 817–70

Charness, Gary, and Angelino Viceisza (2012) 'Comprehension and risk elicitation in the field: evidence from rural Senegal.' *UC–Santa Barbara working paper*

Charness, Gary, and Marie-Claire Villeval (2009) 'Cooperation and competition in intergenerational experiments in the field and in the laboratory.' *American Economic Review* 99(3), 956–78

Chaudhuri, Ananish (2009) *Experiments in Economics: Playing Fair with Money* (New York: Routledge)

Chen, Daniel L., Martin Schonger and Chris Wickens (2016) 'Otree: an open-source platform for laboratory, online, and field experiments.' *Journal of Behavioral and Experimental Finance* 9, 88–97

Chen, Yan, and Tayfun Sönmez (2006) 'School choice: an experimental study.' *Journal of Economic Theory* 127(1), 202–31

Cherry, Todd L., Thomas D. Crocker and Jason F. Shogren (2003) 'Rationality spillovers.' *Journal of Environmental Economics and Management* 45(1), 63–84

Cherry, Todd L., Peter Frykblom and Jason F. Shogren (2002) 'Hardnose the dictator.' *American Economic Review* 92(4), 1218–21

Cherry, Todd L., Peter Frykblom, Jason F. Shogren, John A. List and Melonie Sullivan (2004) 'Laboratory testbeds and non-market valuation: the case of bidding behavior in a second-price auction with an outside option.' *Environmental & Resource Economics* 29(3), 285–94

Chesney, Thomas, Swee-Hoon Chuah and Robert Hoffmann (2009) 'Virtual world experimentation: an exploratory study.' *Journal of Economic Behavior & Organization* 72(1), 618–35

Chetty, Raj (2015) 'Behavioral economics and public policy: a pragmatic perspective.' *American Economic Review* 105(5), 1–33

Chetty, Raj, John N. Friedman, Søren Leth-Petersen, Torben Heien Nielsen and Tore Olsen (2014) 'Active vs. passive decisions and crowd-out in retirement savings accounts: evidence from Denmark.' *Quarterly Journal of Economics* 129(3), 1141–1219

Chetty, Raj, Adam Looney and Kory Kroft (2009) 'Salience and taxation: theory and evidence.' *American Economic Review* 99(4), 1145–77

Choi, Syngjoo, Raymond Fisman, Douglas Gale and Shachar Kariv (2007) 'Consistency and heterogeneity of individual behavior under uncertainty.' *American Economic Review* 97(5), 1921–38

Choi, Syngjoo, Shachar Kariv, Wieland Müller and Dan Silverman (2014) 'Who is (more) rational?' *American Economic Review* 104(6), 1518–50

Choi, James J., David I. Laibson, Brigitte C. Madrian and Andrew Metrick (2004) 'For better or for worse: default effects and 401(k) savings behavior.' In *Perspectives on the Economics of Aging,* ed. David A. Wise (Chicago: The University of Chicago Press), 81–126

Choo, C.Y. Lawrence, Miguel A. Fonseca and Gareth D. Myles (2016) 'Do students behave like real taxpayers in the lab? Evidence from a real effort tax compliance experiment.' *Journal of Economic Behavior & Organization* 124, 102–14

Chou, Eileen, Margaret McConnell, Rosemarie Nagel and Charles R. Plott (2009) 'The control of game form recognition in experiments: understanding dominant strategy failures in a simple two person guessing game.' *Experimental Economics* 12(2), 159–79

Chow, Clare Chua, and Rakesh K. Sarin (2002) 'Known, unknown, and unknowable uncertainties.' *Theory and Decision* 52(2), 127–38

Chu, Yun-Peng, and Ruey-Ling Chu (1990) 'The subsidence of preference reversals in simplified and marketlike experimental settings: a note.' *American Economic Review* 80(4), 902–11

Clemens, Michael A. (2016) 'The meaning of failed replications: a review and proposal.' *Journal of Economic Surveys*, forthcoming

Coase, Ronald H. (1960) 'The problem of social cost.' *Journal of Law & Economics* 3, 1–44

Coble, Keith H., and Jayson L. Lusk (2010) 'At the nexus of risk and time preferences: an experimental investigation.' *Journal of Risk and Uncertainty* 41(1), 67–79

Coffman, Lucas C., and Muriel Niederle (2015) 'Pre-analysis plans have limited upside, especially where replications are feasible.' *Journal of Economic Perspectives* 29(3), 81–98

Cohen, Jonathan D., Keith Marzilli Ericson, David I. Laibson and John Myles White (2016) 'Measuring time preferences.' *NBER working paper*

Cohen, Michele, Jean-Yves Jaffray and Tanios Said (1987) 'Experimental comparison of individual behavior under risk and under uncertainty for gains and for losses.' *Organizational Behavior and Human Decision Processes* 39(1), 1–22

Coller, Maribeth, and Melonie B. Williams (1999) 'Eliciting individual discount rates.' *Experimental Economics* 2(2), 107–27

Conte, Anna, John D. Hey and Peter G. Moffatt (2011) 'Mixture models of choice under risk.' *Journal of Econometrics* 162(1), 79–88

Cook, Diane J., and Wenzhan Song (2009) 'Ambient intelligence and wearable computing: sensors on the body, in the home, and beyond.' *Journal of Ambient Intelligence and Smart Environments* 1(2), 83–6

Cooper, David J., and Hanming Fang (2008) 'Understanding overbidding in second price auctions: an experimental study.' *Economic Journal* 118(532), 1572–95

Cooper, David J., and John H. Kagel (2003) 'The impact of meaningful context on strategic play in signaling games.' *Journal of Economic Behavior & Organization* 50(3), 311–37

—— (2009) 'The role of context and team play in cross-game learning.' *Journal of the European Economic Association* 7(5), 1101–39

Cooper, Russell W., Douglas V. DeJong, Robert Forsythe and Thomas W. Ross (1990) 'Selection criteria in coordination games: some experimental results.' *American Economic Review* 80(1), 218–33

—— (1996) 'Cooperation without reputation: experimental evidence from prisoner's dilemma games.' *Games and Economic Behavior* 12(2), 187–218

Costa-Gomes, Miguel A., and Vincent P. Crawford (2006) 'Cognition and behavior in two-person guessing games: an experimental study.' *American Economic Review* 96(5), 1737–68

Costa-Gomes, Miguel A., and Georg Weizsäcker (2008) 'Stated beliefs and play in normal-form games.' *Review of Economic Studies* 75(3), 729–62

Cox, James C., Vjollca Sadiraj and Ulrich Schmidt (2015) 'Paradoxes and mechanisms for choice under risk.' *Experimental Economics* 18(2), 215–50

Crawford, Vincent P., Miguel A. Costa-Gomes and Nagore Iriberri (2013) 'Structural models of nonequilibrium strategic thinking: theory, evidence, and applications.' *Journal of Economic Literature* 51(1), 5–62

Crawford, Vincent P., and Nagore Iriberri (2007a) 'Fatal attraction: salience, naïveté, and sophistication in experimental "hide-and-seek" games.' *American Economic Review* 97(5), 1731–50

—— (2007b) 'Level-*k* auctions: can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?' *Econometrica* 75(6), 1721–70

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot and Philippe Zamora (2013) 'Do labor market policies have displacement effects? Evidence from a clustered randomized experiment.' *Quarterly Journal of Economics* 128(2), 531–80

Cronqvist, Henrik, and Richard H. Thaler (2004) 'Design choices in privatized social-security systems: learning from the Swedish experience.' *American Economic Review* 94(2), 424–8

Crosetto, Paolo, and Antonio Filippin (2016a) 'A theoretical and experimental appraisal of four risk elicitation methods.' *Experimental Economics* 19(3), 613–41

—— (2016b) 'A theoretical and experimental appraisal of four risk elicitation methods.' *Experimental Economics* 19, 613–641

Croson, Rachel T. (2000) 'Thinking like a game theorist: factors affecting the frequency of equilibrium play.' *Journal of Economic Behavior & Organization* 41(3), 299–314

—— (2006) 'Contrasting methods and comparative findings in psychology and economics.' In *Social Psychology and Economics,* ed. David De Cremer, Marcel Zeelenberg, and J. Keith Murnighan (New York: Lawrence Erlbaum Associates)

Croson, Rachel T., and Simon Gächter (2010) 'The science of experimental economics.' *Journal of Economic Behavior & Organization* 73(1), 122–31

Crump, Matthew J.C., John V. McDonnell and Todd M. Gureckis (2013) 'Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research.' *PLoS ONE* 8(3), e57410

Cubitt, Robin P., Michalis Drouvelis, Simon Gächter and Ruslan Kabalin (2011) 'Moral judgments in social dilemmas: how bad is free riding?' *Journal of Public Economics* 95(3), 253–64

Cubitt, Robin P., and Daniel Read (2007) 'Can intertemporal choice experiments elicit time preferences for consumption?' *Experimental Economics* 10(4), 369–89

Cubitt, Robin P., Chris Starmer and Robert Sugden (1998) 'On the validity of the random lottery incentive system.' *Experimental Economics* 1(2), 115–31

Cummings, Ronald G., Glenn W. Harrison and Laura Osborne (1995) 'Can the bias of contingent valuation be reduced? Evidence from the laboratory.' *University of South Carolina, economics working paper*

Cummings, Ronald G., and Laura O. Taylor (1998) 'Does realism matter in contingent valuation surveys?' *Land Economics* 74(2), 203–15

—— (1999) 'Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method.' *American Economic Review* 89(3), 649–65

Curley, Shawn P., and Frank J. Yates (1989) 'An empirical evaluation of descriptive models of ambiguity reactions in choice situations.' *Journal of Mathematical Psychology* 33(4), 397–427

Danz, David N., Dietmar Fehr and Dorothea Kübler (2012) 'Information and beliefs in a repeated normal-form game.' *Experimental Economics* 15(4), 622–40

Davidson, Russell, and James G. MacKinnon (2004) *Econometric Theory and Methods* (New York: Oxford University Press)

De Finetti, Bruno (1974) *Theory of Probability*, vols. 1–2 (New York: Wiley)

Deaton, Angus (2010) 'Instruments, randomization, and learning about development.' *Journal of Economic Literature* 48(2), 424–55

DellaVigna, Stefano (2009) 'Psychology and economics: evidence from the field.' *Journal of Economic Literature* 47(2), 315–72

DellaVigna, Stefano, and Ulrike Malmendier (2006) 'Paying not to go to the gym.' *American Economic Review* 96(3), 694–719

Delquié, Philippe (1993) 'Inconsistent trade-offs between attributes: new evidence in preference assessment biases.' *Management Science* 39(11), 1382–95

Denant-Boèmont, Laurent, Enrico Diecidue and Olivier L'Haridon (2017) 'Patience and time consistency in collective decisions.' *Experimental Economics* 20(1), 181–208

Devetag, Giovanna, and Andreas Ortmann (2007) 'When and why? A critical survey on coordination failure in the laboratory.' *Experimental Economics* 10(3), 331–44

Diamond, Peter A., and Jerry A. Hausman (1994) 'Contingent valuation: is some number better than no number?' *Journal of Economic Perspectives* 8(4), 45–64

Diecidue, Enrico, and Peter P. Wakker (2001) 'On the intuition of rank-dependent utility.' *Journal of Risk and Uncertainty* 23(3), 281–98

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp and Gert G. Wagner (2011) 'Individual risk attitudes: measurement, determinants, and behavioral consequences.' *Journal of the European Economic Association* 9(3), 522–50

Dominitz, Jeff, and Charles F. Manski (1997) 'Using expectations data to study subjective income expectations.' *Journal of the American Statistical Association* 92(439), 855–67

Dougherty, Christopher (2006) *Introduction to Econometrics* (New York: Oxford University Press)

Drasgow, Fritz (1988) *Polychoric and polyserial correlations* (Wiley Online Library)

Drichoutis, Andreas, and Jayson L. Lusk (2016) 'What can multiple price lists really tell us about risk preferences?' *Journal of Risk and Uncertainty* 53(2–3), 89–106

Dufwenberg, Martin, and Georg Kirchsteiger (2004) 'A theory of sequential reciprocity.' *Games and Economic Behavior* 47(2), 268–98

Dutcher, Glenn, Tim Salmon and Krista Jabs Saral (2015) 'Is "real" effort more real?' *Working paper*

Duvendack, Maren, Richard W. Palmer-Jones and Robert W. Reed (2015) 'Replications in economics: a progress report.' *Econ Journal Watch* 12(2), 164–91

Dyer, Douglas, John H. Kagel and Dan Levin (1989) 'A comparison of naive and experienced bidders in common value offer auctions: a laboratory analysis.' *Economic Journal* 99(394), 108–15

Ebert, Jane E.J., and Drazen Prelec (2007) 'The fragility of time: time-insensitivity and valuation of the near and far future.' *Management Science* 53(9), 1423–38

Echenique, Federico, Alistair J. Wilson and Leeat Yariv (2016) 'Clearinghouses for two-sided matching: an experimental study.' *Quantitative Economics* 7(2), 449–82

Echenique, Federico, and Leeat Yariv (2013) 'An experimental study of decentralized matching.' *CalTech working paper*

Eckel, Catherine C., and Philip J. Grossman (2008) 'Forecasting risk attitudes: An experimental study using actual and forecast gamble choices.' *Journal of Economic Behavior & Organization* 68, 1–17

Eckel, Catherine C., Philip J. Grossman and Rachel M. Johnston (2005) 'An experimental test of the crowding out hypothesis.' *Journal of Public Economics* 89(8), 1543–60

Eeckhoudt, Louis, Christian Gollier and Harris Schlesinger (2005) *Economic and Financial Decisions under Risk* (Princeton, NJ: Princeton University Press)

Ellsberg, Daniel (1961) 'Risk, ambiguity, and the savage axioms.' *Quarterly Journal of Economics* 75(4), 643–69

Engel, Christoph (2011) 'Dictator games: a meta study.' *Experimental Economics* 14(4), 583–610

Epley, Nicholas, and Thomas Gilovich (2006) 'The anchoring-and-adjustment heuristic why the adjustments are insufficient.' *Psychological Science* 17(4), 311–18

Epper, Thomas, Helga Fehr-Duda and Adrian Bruhin (2011) 'Viewing the future through a warped lens: why uncertainty generates hyperbolic discounting.' *Journal of Risk and Uncertainty* 43(3), 169–203

Eriksson, Tor, Sabrina Teyssier and Marie-Claire Villeval (2009) 'Self-selection and the efficiency of tournaments.' *Economic Inquiry* 47(3), 530–48

Etchart-Vincent, Nathalie (2004) 'Is probability weighting sensitive to the magnitude of consequences? An experimental investigation on losses.' *Journal of Risk and Uncertainty* 28(3), 217–35

Evans, Anthony M., Kyle D. Dillon and David G. Rand (2015) 'Fast but not intuitive, slow but not reflective: decision conflict drives reaction times in social dilemmas.' *Journal of Experimental Psychology: General* 144(5), 951–66

Falk, Armin (2007) 'Gift exchange in the field.' *Econometrica* 75(5), 1501–11

Falk, Armin, and Ernst Fehr (2003) 'Why labour market experiments?' *Labour Economics* 10(4), 399–406

Falk, Armin, Ernst Fehr and Urs Fischbacher (2003) 'On the nature of fair behavior.' *Economic Inquiry* 41(1), 20–6

Falk, Armin, and James J. Heckman (2009) 'Lab experiments are a major source of knowledge in the social sciences.' *Science* 326(5952), 535–8

Farquhar, Peter (1984) 'Utility assessment methods.' *Management Science* 30(11), 1283–1300

Featherstone, Clayton, and Muriel Niederle (2016) 'Improving on strategy-proof school choice mechanisms: an experimental investigation.' *Games and Economic Behavior* 100, 353–75

Fechner, Gustav T. (1869) *Elemente der psychophysik* (Leipzig: Breitkopf und Härtel)

Fehr, Ernst, Georg Kirchsteiger and Arno Riedl (1993) 'Does fairness prevent market clearing? An experimental investigation.' *Quarterly Journal of Economics* 108(2), 437–59

Fehr, Ernst, and Klaus M. Schmidt (1999) 'A theory of fairness, competition, and cooperation.' *Quarterly Journal of Economics* 114(3), 817–68

—— (2010) 'On inequity aversion: a reply to binmore and Shaked.' *Journal of Economic Behavior & Organization* 73(1), 101–8

Fehr-Duda, Helga, and Thomas Epper (2012) 'Probability and risk: foundations and economic implications of probability-dependent risk preferences.' *Annual Review of Economics* 4(1), 567–93

Fennema, Hein, and Marcel van Assen (1998) 'Measuring the utility of losses by means of the tradeoff method.' *Journal of Risk and Uncertainty* 17(3), 277–96

Ferraro, Paul J., and Christian A. Vossler (2010) 'The source and significance of confusion in public goods experiments.' *B.E. Journal of Economic Analysis & Policy* 10(1), 1–42

Fey, Mark, Richard D. McKelvey and Thomas R. Palfrey (1996) 'An experimental study of constant-sum centipede games.' *International Journal of Game Theory* 25(3), 269–87

Fiedler, Marina, and Ernan Haruvy (2009) 'The lab versus the virtual lab and virtual field: an experimental investigation of trust games with communication.' *Journal of Economic Behavior & Organization* 72(2), 716–24

Fiorina, Morris P., and Charles R. Plott (1978) 'Committee decisions under majority rule: an experimental study.' *American Political Science Review* 72(2), 575–98

Fischbacher, Urs (2007) 'z-tree: Zurich toolbox for ready-made economic experiments.' *Experimental Economics* 10(2), 171–8

Fishburn, Peter C. (1989) 'Retrospective on the utility theory of Von Neumann and Morgenstern.' *Journal of Risk and Uncertainty* 2(2), 127–57

Fishburn, Peter C., and Ariel Rubinstein (1982) 'Time preference.' *International Economic Review* 23(3), 677–94

Fontaine, Philippe, and Robert Leonard (2005) *The Experiment in the History of Economics* (London: Routledge)

Forsythe, Robert, Joel L. Horowitz, N. Eugene Savin and Martin Sefton (1994) 'Fairness in simple bargaining experiments.' *Games and Economic Behavior* 6(3), 347–69

Fox, Craig R., and Russel A. Poldrack (2014) 'Prospect theory and the brain.' In *Handbook of Neuroeconomics,* ed. P. Glimcher and E. Fehr (New York: Elsevier), 533–68

Fox, John A., Jason F. Shogren, Dermot J. Hayes and James B. Kliebenstein (1998) 'Cvm-x: calibrating contingent values with experimental auction markets.' *American Journal of Agricultural Economics* 80(3), 455–65

Fox, Craig R., and Amos Tversky (1995) 'Ambiguity aversion and comparative ignorance.' *Quarterly Journal of Economics* 110(3), 585–603

Frank, Bjorn (1998) 'Good news for experimenters: subjects do not care about your welfare.' *Economics Letters* 61(2), 171–4

Frank, Bjorn, and Gunther G. Schulze (2000) 'Does economics make citizens corrupt?' *Journal of Economic Behavior & Organization* 43(1), 101–13

Fréchette, Guillaume R. (2015) 'Laboratory experiments: professionals versus students.' In *Handbook of Experimental Economic Methodology,* ed. Guillaume Fréchette and Andrew Schotter (New York: Oxford University Press), 360–90

—— (2016) 'Experimental economics across subject population.' In *Handbook of Experimental Economic,* ed. John H. Kagel and Alvin E. Roth, vol. 2 (Princeton, NJ: Princeton University Press), 435–80

Fréchette, Guillaume R., and Andrew Schotter, eds. (2015) *Handbook of Experimental Economic Methodology* (New York: Oxford University Press)

Frederick, Shane (2005) 'Cognitive reflection and decision making.' *Journal of Economic Perspectives* 19(4), 25–42

Frederick, Shane, and George Loewenstein (2008) 'Conflicting motives in evaluations of sequences.' *Journal of Risk and Uncertainty* 37(2–3), 221–35

Frederick, Shane, George Loewenstein and Ted O'donoghue (2002) 'Time discounting and time preference: a critical review.' *Journal of Economic Literature* 40(2), 351–401

Frey, Bruno S., and Reto Jegen (2001) 'Motivation crowding theory.' *Journal of Economic Surveys* 15(5), 589–611

Friedman, Daniel, and Dominic W. Massaro (1998) 'Understanding variability in binary and continuous choice.' *Psychonomic Bulletin & Review* 5(3), 370–89

Friedman, Daniel, and Shyam Sunder (1994) *Experimental Methods: A Primer for Economists* (Cambridge: Cambridge University Press)

Friedman, Milton (1953) *Essays in Positive Economics* (Chicago: The University of Chicago Press)

Fudenberg, Drew, and David K. Levine (2006) 'A dual-self model of impulse control.' *American Economic Review* 96(5), 1449–76

Fudenberg, Drew, David K. Levine and Zacharias Maniadis (2012) 'On the robustness of anchoring effects in WTP and WTA experiments.' *American Economic Journal: Microeconomics* 4(2), 131–45

Gächter, Simon, Eric J. Johnson and Andreas Herrmann (2007) 'Individual-level loss aversion in riskless and risky choices.' *IZA discussion paper*

Gächter, Simon, and Elke Renner (2010) 'The effects of (incentivized) belief elicitation in public goods experiments.' *Experimental Economics* 13(3), 364–77

Gächter, Simon, Chris Starmer and Fabio Tufano (2015a) 'Measuring the closeness of relationships: a comprehensive evaluation of the "inclusion of the other in the self' scale." *PLoS ONE* 10(6), e0129478

—— (2015b) 'Measuring the impact of social relationships: the value of 'oneness'.' *Working paper*

Gajdos, Thibault, Takashi Hayashi, Jean-Marc Tallon and Jean-Christophe Vergnaud (2008) 'Attitude toward imprecise information.' *Journal of Economic Theory* 140(1), 27–65

Geanakoplos, John, David Pearce and Ennio Stacchetti (1989) 'Psychological games and sequential rationality.' *Games and Economic Behavior* 1(1), 60–79

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin (2014) *Bayesian Data Analysis*, vol. 2 (Boca Raton: Taylor & Francis)

Gerber, Alan S., and Donald P. Green (2012) *Field Experiments: Design, Analysis, and Interpretation* (New York: W. W. Norton & Co.)

Ghirardato, Paolo, Fabio Maccheroni and Massimo Marinacci (2004) 'Differentiating ambiguity and ambiguity attitude.' *Journal of Economic Theory* 118(2), 133–73

Ghirardato, Paolo, and Massimo Marinacci (2001) 'Risk, ambiguity, and the separation of utility and beliefs.' *Mathematics of Operations Research* 26(4), 864–90

Gibbons, Jean Dickinson (2011) *Nonparametric Statistical Inference* (Boca Raton: Chapman and Hall/CRC)

Gilboa, Itzhak (2009) *Theory of Decision under Uncertainty* (New York: Cambridge University Press)

Gilboa, Itzhak, and David Schmeidler (1989) 'Maxmin expected utility with non-unique prior.' *Journal of Mathematical Economics* 18(2), 141–53

Gill, David, and Victoria L. Prowse (2015) 'A novel computerized real effort task based on sliders.' *Working paper*

Gino, Francesca, and Cassie Mogilner (2014) 'Time, money, and morality.' *Psychological Science* 25(2), 414–21

Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman and Christine L. Soutter (2000) 'Measuring trust.' *Quarterly Journal of Economics* 115(3), 811–46

Gneezy, Uri, and John A. List (2006) 'Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments.' *Econometrica* 74(5), 1365–84

Gneezy, Uri, Muriel Niederle and Aldo Rustichini (2003) 'Performance in competitive environments: gender differences.' *Quarterly Journal of Economics* 118(3), 1049–74

Gneezy, Uri, and Jan Potters (1997) 'An experiment on risk taking and evaluation periods.' *Quarterly Journal of Economics* 112(2), 631–45

Gneezy, Uri, and Aldo Rustichini (2000) 'Pay enough or don't pay at all.' *Quarterly Journal of Economics* 115(3), 791–810

Goeree, Jacob K., Charles A. Holt and Thomas R. Palfrey (2002) 'Quantal response equilibrium and overbidding in private-value auctions.' *Journal of Economic Theory* 104(1), 247–72

—— (2005) 'Regular quantal response equilibrium.' *Experimental Economics* 8(4), 347–67

Goeschl, Timo, Sara Elisa Kettner, Johannes Lohse and Christiane Schwieren (2017) 'What do we learn from public good games about voluntary climate action? Evidence from an artefactual field experiment.' *Environmental & Resource Economics*, forthcoming

Goldin, Jacob, and Daniel Reck (2015) 'Preference identification under inconsistent choice.' *Princeton University working paper*

Good, Irving John (1952) 'Rational decisions.' *Journal of the Royal Statistical Society. Series B (Methodological)* 14(1), 107–14

Green, Donald P., Karen E. Jacowitz, Daniel Kahneman and Daniel McFadden (1998) 'Referendum contingent valuation, anchoring, and willingness to pay for public goods.' *Resource and Energy Economics* 20(2), 85–116

Greenberg, David, and Burt S. Barnow (2014) 'Flaws in evaluations of social programs: illustrations from randomized controlled trials.' *Evaluation Review* 38(5), 359–87

Greiner, Ben (2015) 'Subject pool recruitment procedures: organizing experiments with orsee.' *Journal of the Economic Science Association* 1(1), 114–25

Grether, David M., and Charles R. Plott (1979) 'Economic theory of choice and the preference reversal phenomenon.' *American Economic Review* 69(4), 623–38

—— (1984) 'The effects of market prices in oligopolistic markets: an experimental examination of the ethyl case.' *Economic Inquiry* 22(4), 479–507

Guala, Francesco (2002) 'On the scope of experiments in economics: comments on siakantaris.' *Cambridge Journal of Economics* 26(2), 261–7

—— (2005) *The Methodology of Experimental Economics* (Cambridge: Cambridge University Press)

Guala, Francesco, and Luigi Mittone (2005) 'experiments in economics: external validity and the robustness of phenomena.' *Journal of Economic Methodology* 12(4), 495–515

Guth, Werner, Rolf Schmittberger and Bernd Schwarze (1982) 'An experimental analysis of ultimatum bargaining.' *Journal of Economic Behavior & Organization* 3(4), 367–88

Haile, Philip A., Ali Hortaçsu and Grigory Kosenok (2008) 'On the empirical content of quantal response equilibrium.' *American Economic Review* 98(1), 180–200

Halevy, Yoram (2007) 'Ellsberg revisited: an experimental study.' *Econometrica* 75(2), 503–36

—— (2015) 'Time consistency: stationarity and time invariance.' *Econometrica* 83(1), 335–52

Hall, Joshua (2005) 'Homer economicus: using *The Simpsons* to teach economics.' *Journal of Private Enterprise* 20(2), 166–77

—— (2014) *Homer Economicus:* The Simpsons and Economics (Stanford, CA: Stanford Economics and Finance)

Hammond, Kenneth R. (1998) 'Ecological validity: then and now.' *Brunswik Society Notes and Essays*, www.brunswik.org/notes/essay2.html

Hammond, Peter, and Horst Zank (2014) 'Handbook of the economics of risk and uncertainty.' In *Rationality and Dynamic Consistency under Risk and Uncertainty,* ed. Mark J. Machina and W. Kip Viscusi (Amsterdam: Elsevier), 41–97

Hanson, Robin, Ryan Oprea and David Porter (2006) 'Information aggregation and manipulation in an experimental market.' *Journal of Economic Behavior & Organization* 60(4), 449–59

Hao, Li, and Daniel Houser (2012) 'Belief elicitation in the presence of naïve respondents: an experimental study.' *Journal of Risk and Uncertainty* 44(2), 161–80

Harless, David W., and Colin F. Camerer (1994) 'The predictive utility of generalized expected utility theories.' *Econometrica* 62(6), 1251–89

Harrison, Glenn W., Eric J. Johnson, Melayne M. McInnes and Elisabet E. Rutström (2005) 'Risk aversion and incentive effects: comment.' *American Economic Review* 95(3), 897–901

Harrison, Glenn W., and John A. List (2004) 'Field experiments.' *Journal of Economic Literature* 42(4), 1009–55

Harrison, Glenn W., Jimmy Martínez-Correa and J. Todd Swarthout (2014) 'Eliciting subjective probabilities with binary lotteries.' *Journal of Economic Behavior & Organization* 101, 128–40

Harrison, Glenn W., and Elisabet E. Rutström (2008) 'Experimental evidence on the existence of hypothetical bias in value elicitation methods.' In *Handbook of Experimental Economics Results,* ed. Charles R. Plott and Vernon L. Smith, vol. 1 (Amsterdam: Elsevier) pp. 752–67

Haruvy, Ernan, and M. Utku Ünver (2007) 'Equilibrium selection and the role of information in repeated matching markets.' *Economics Letters* 94(2), 284–9

Harvey, Charles M. (1986) 'Value functions for infinite-period planning.' *Management Science* 32(9), 1123–39

Häubl, Gerald, Benedict G.C. Dellaert and Bas Donkers (2010) 'Tunnel vision: local behavioral influences on consumer decisions in product search.' *Marketing Science* 29(3), 438–55

Healy, Paul J., Sera Linardi, Richard J. Lowery and John O. Ledyard (2010) 'Prediction markets: alternative mechanisms for complex environments with few traders.' *Management Science* 56(11), 1977–96

Heckman, James J. (1996) 'Identification of causal effects using instrumental variables: comment.' *Journal of the American Statistical Association* 91(434), 459–62

—— (1997) 'Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations.' *Journal of Human Resources* 32(3), 441–62

—— (2010) 'Building bridges between structural and program evaluation approaches to evaluating policy.' *Journal of Economic Literature* 48(2), 356–98

Heckman, James J., Robert J. Lalonde and Jeffrey A. Smith (1999) 'The economics and econometrics of active labor market programs.' In *Handbook of Labor Economics,* ed. Orley C. Ashenfelter and David Card, vol. 3, Part 1 (Amsterdam: Elsevier Science) pp. 1865–2097

Hedges, Larry V., and Ingram Olkin (1985) *Statistical Methods for Meta-Analysis* (San Diego: Academic Press)

Heinemann, Frank, Rosemarie Nagel and Peter Ockenfels (2009) 'Measuring strategic uncertainty in coordination games.' *Review of Economic Studies* 76(1), 181–221

Henrich, Joseph, Steve J. Heine and Ara Norenzayan (2010) 'The weirdest people in the world?' *Behavioral and Brain Sciences* 33(2–3), 61–83

Hergueux, Jérome, and Nicolas Jacquemet (2015) 'Social preferences in the online laboratory: a randomized experiment.' *Experimental Economics* 18(2), 252–83

Herrnstein, Richard J. (1981) 'Self-control as response strength.' In *The Matching Law: Papers in Psychology and Economics,* ed. Richard J. Herrnstein, Howard Rachlin and David I. Laibson (Cambridge, MA: Harvard Universtity Press) pp. 3–20

Hershey, John C., Howard Kunreuther and Paul J.H. Schoemaker (1982) 'Sources of bias in assesment procedure for utility functions.' *Management Science* 28(8), 936–54

Hershey, John C., and Paul J.H. Schoemaker (1985) 'Probability versus certainty equivalence methods in utility measurement: are they equivalent?' *Management Science* 31(10), 1213–31

Hertwig, Ralph, and Andreas Ortmann (2001) 'Experimental practices in economics: a methodological challenge for psychologists?' *Behavioral and Brain Sciences* 24(3), 383–403

Hey, John D. (2005) 'Why we should not be silent about noise.' *Experimental Economics* 8(4), 325–45

—— (2014) 'Choice under uncertainty: empirical methods and experimental results.' In *Handbook of the Economics of Risk and Uncertainty,* ed. Mark J. Machina and Kip Viscusi, vol. 1 of *Handbook of the Economics of Risk and Uncertainty* (Amsterdam: North Holland) pp. 809–50

Hey, John D., and Jinkwon Lee (2005) 'Do subjects separate (or are they sophisticated)?' *Experimental Economics* 8(3), 233–65

Hey, John D., Andrea Morone and Ulrich Schmidt (2009) 'Noise and bias in eliciting preferences.' *Journal of Risk and Uncertainty* 39(3), 213–35

Hey, John D., and Chris Orme (1994) 'Investigating generalizations of expected utility theory using experimental data.' *Econometrica* 62(6), 1391–1426

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat and Vernon L. Smith (1994) 'Preferences, property rights, and anonymity in bargaining games.' *Games and Economic Behavior* 7(3), 346–80

Hoffman, Elizabeth, Kevin McCabe and Vernon L. Smith (1996) 'Social distance and other-regarding behavior in dictator games.' *American Economic Review* 86(3), 653–60

Holcomb, James H., and Paul S. Nelson (1992) 'Another experimental look at individual time preference.' *Rationality and Society* 4(2), 199–220

Holt, Charles A. (1986) 'Scoring-rule procedures for eliciting subjective probability and utility functions.' In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti,* ed. Prem K. Goel and A. Zellner (Amsterdam: North Holland Press), 279–90

—— (2006) *Markets, Games, & Strategic Behavior* (Boston, MA: Addison Wesley)

Holt, Charles A., and Monica Capra (2000) 'Classroom games: a prisoner's dilemma.' *Journal of Economic Education* 31(3), 229–36

Holt, Charles A., and Susan K. Laury (1997) 'Classroom games: voluntary provision of a public good.' *Journal of Economic Perspectives* 11(4), 209–15

Holt, Charles A., and David T. Scheffman (1987) 'Facilitating practices: the effects of advance notice and best-price policies.' *RAND Journal of Economics* 18(2), 187–97

Holt, Charles A., and Roger Sherman (2014) 'Risk aversion and the winner's curse.' *Southern Economic Journal* 81(1), 7–22

—— (2002) 'Risk aversion and incentive effects.' *American Economic Review* 92(4), 1644–55

Holt, Charles A., William M. Shobe and Angela M. Smith (2006) 'An experimental basis for public policy initiatives.' In *Promoting the General Welfare: American Democracy and the Political Economy of Government Performance,* ed. Alan S. Gerber and Eric M. Patashnik (Washington, DC: Brookings Institution Press), 174–96

Horton, John J., David G. Rand and Richard J. Zeckhauser (2011) 'The online laboratory: conducting experiments in a real labor market.' *Experimental Economics* 14(3), 399–425

Hossain, Tanjim, and Ryo Okui (2013) 'The binarized scoring rule.' *Review of Economic Studies* 80(3), 984–1001

Huck, Steffen, Hans-Theo Normann and Jorg Oechssler (2004) 'Two are few and four are many: number effects in experimental oligopolies.' *Journal of Economic Behavior & Organization* 53(4), 435–46

Huck, Steffen, and Georg Weizsäcker (2002) 'Do players correctly estimate what others do? Evidence of conservatism in beliefs.' *Journal of Economic Behavior & Organization* 47(1), 71–85

Hyndman, Kyle B., Antoine Terracol and Jonathan Vaksmann (2013) 'Beliefs and (in)stability in normal-form games.' *Working paper*

Ioannidis, John P. A. (2005) 'Why most published research findings are false.' *PLoS Med* 2(8), e124

Ioannidis, John P. A., and Chris Doucouliagos (2013) 'What's to know about the credibility of empirical economics?' *Journal of Economic Surveys* 27(5), 997–1004

Isaac, Mark R., James M. Walker and Susan H. Thomas (1984) 'Divergent evidence on free riding: an experimental examination of possible explanations.' *Public Choice* 43(2), 113–49

Isaac, Mark R., James M. Walker and Arlington W. Williams (1994) 'Group size and the voluntary provision of public goods: experimental evidence utilizing large groups.' *Journal of Public Economics* 54(1), 1–36

Isoni, Andrea, Graham Loomes and Robert Sugden (2011) 'The willingness to pay–willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: comment.' *American Economic Review* 101(2), 991–1011

Iyengar, Sheena S, Rachael E. Wells and Barry Schwartz (2006) 'Doing better but feeling worse: looking for the "best" job undermines satisfaction.' *Psychological Science* 17(2), 143–50

Jacquemet, Nicolas, Robert-Vincent Joule, Stéphane Luchini and Jason F. Shogren (2009) 'Earned wealth, engaged bidders? Evidence from a second price auction.' *Economics Letters* 105(1), 36–8

—— (2011) 'Do people always pay less than they say? Testbed laboratory experiments with IV and HG values.' *Journal of Public Economic Theory* 13(5), 857–82

Jacquemet, Nicolas, and Frédéric Koessler (2013) 'Using or hiding private information? An experimental study of zero-sum repeated games with incomplete information.' *Games and Economic Behavior* 78, 103–20

Jacquemet, Nicolas, Stéphane Luchini, Julie Rosaz and Jason F. Shogren (2017) 'Truth-telling under oath.' *Management Science*, forthcoming

Jensen, Floyd A., and Cameron R. Peterson (1973) 'Psychological effects of proper scoring rules.' *Organizational Behavior and Human Performance* 9(2), 307–17

Jensen, Niels Erik (1967) 'An introduction to Bernoullian utility theory: I. utility functions.' *Swedish Journal of Economics,* 163–83

Johnson, Eric J., and Daniel G. Goldstein (2003) 'Do defaults save lives?' *Science* 302(5649), 1338–39

Johnson, Eric J., John Hershey, Jacqueline Meszaros and Howard Kunreuther (1993) 'Framing, probability distortions, and insurance decisions.' *Journal of Risk and Uncertainty* 7(1), 35–51

Johnson, Eric J., and David A. Schkade (1989) 'Bias in utility assessments: further evidence and explanations.' *Management Science* 35(4), 406–24

Johnson, Eric J., Suzanne B. Shu, Benedict G.C. Dellaert, Craig R. Fox, Daniel G. Goldstein, Gerald Häubl, Richard P. Larrick, John W. Payne, Ellen Peters, David A. Schkade, Brian Wansink and Elke U. Weber (2012) 'Beyond nudges: tools of a choice architecture.' *Marketing Letters* 23(2), 487–504

Johnson, Martin A. (1984) 'Concern for appropriateness scale and behavioral conformity.' *Journal of Personality Assessment* 53(3), 567–74

Johnson, Noel D., and Alexandra Mislin (2012) 'How much should we trust the World Values Survey trust question?' *Economics Letters* 116(2), 210–12

Kadane, Joseph B., and Robert L. Winkler (1988) 'Separating probability elicitation from utilities.' *Journal of the American Statistical Association* 83(402), 357–63

Kagel, John H., and Dan Levin (1986) 'The winner's curse and public information in common value auctions.' *American Economic Review* 76(5), 894–920

Kagel, John H., and Alvin E. Roth (1995) *Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press)

—— (2000) 'The dynamics of reorganization in matching markets: a laboratory experiment motivated by a natural experiment.' *Quarterly Journal of Economics* 115(1), 201–35

Kahneman, Daniel (1988) 'Experimental economics: a psychological perspective.' In *Bounded Rational Behavior in Experimental Games and Markets* ed. Reinhard Tietz, Wulf Albers and Reinhard Selten (Berlin: Springer), 11–18

——(2003) 'Maps of bounded rationality: Psychology for behavioral economics.' *American Economic Review* 93(5), 1449–75
——(2011) *Thinking, Fast and Slow* (New York: Macmillan)
——(2012) 'A proposal to deal with questions about priming effects.' *Open letter, available at* www.nature.com/polopoly_fs/7.6716.1349271308!/suppinfoFile/KahnemanLetter.pdf
Kahneman, Daniel, Jack L. Knetsch and Richard H. Thaler (1986) 'Fairness and the assumptions of economics.' *Journal of Business* 59(4), S285–300
——(1990) 'Experimental tests of the endowment effect and the Coase theorem.' *Journal of Political Economy* 98(6), 1325–48
Kahneman, Daniel, and Amos Tversky (1979) 'Prospect theory: an analysis of decision under risk.' *Econometrica* 47(2), 263–91
Kahneman, Daniel, Peter P. Wakker and Rakesh K. Sarin (1997) 'Back to Bentham? Explorations of experienced utility.' *Quarterly Journal of Economics* 112(2), 375–406
Karmarkar, Uday S. (1978) 'Subjectively weighted utility: a descriptive extension of the expected utility model.' *Organizational Behavior and Human Performance* 21(1), 61–72
Karni, Edi (2009) 'A mechanism for eliciting probabilities.' *Econometrica* 77(2), 603–6
Kass, Robert E., and Adrian E. Raftery (1995) 'Bayes factors.' *Journal of the American Statistical Association* 90(430), 773–95
Kawagoe, Toshiji, and Hirokazu Takizawa (2012) 'Level-*k* analysis of experimental centipede games.' *Journal of Economic Behavior & Organization* 82(2–3), 548–66
Keane, Michael P. (2010) 'Structural vs. atheoretic approaches to econometrics.' *Journal of Econometrics* 156(1), 3–20
Kessler, Judd B., and Lise Vesterlund (2015) 'The external validity of laboratory experiments: the misleading emphasis on quantitative effects.' In *Handbook of Experimental Economic Methodology,* ed. Guillaume Fréchette and Andrew Schotter (Oxford: Oxford University Press)
Keynes, John Maynard (1936) *The General Theory of Employment, Interest, and Money* (London: Macmillan)
Kirby, Kris N., and Nino N. Maraković (1995) 'Modeling myopic decisions: evidence for hyperbolic delay-discounting within subjects and amounts.' *Organizational Behavior and Human Decision Processes* 64(1), 22–30
Kirby, Kris N., and Mariana Santiesteban (2003) 'Concave utility, transaction costs, and risk in measuring discounting of delayed rewards.' *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(1), 66–79
Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Stepan Bahnik, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, Jane S. Hunt, Jeffrey R. Huntsinger, Hans Ijzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. van Swol, Donna Thompson, Anna E. van't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka and Brian A. Nosek (2014) 'Investigating variation in replicability: a "many labs" replication project.' *Social Psychology* 45(3), 142–52
Klibanoff, Peter, Massimo Marinacci and Sujoy Mukerji (2005) 'A smooth model of decision making under ambiguity.' *Econometrica* 73(6), 1849–92

Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee Vermeulen and Marian V. Wrobel (2012) 'Comparison friction: experimental evidence from Medicare drug plans.' *Quarterly Journal of Economics* 127(1), 199–235

Knetsch, Jack L. (1989) 'The endowment effect and evidence of nonreversible indifference curves.' *American Economic Review* 79(5), 1277–84

Köbberling, Veronika, and Peter P. Wakker (2005) 'An index of loss aversion.' *Journal of Economic Theory* 122(1), 119–31

Koehler, Derek J. (1991) 'Explanation, imagination, and confidence in judgment.' *Psychological Bulletin* 110(3), 499–519

Kotchen, Matthew J., and Stephen D. Reiling (1999) 'Do reminders of substitutes and budget constraints influence contingent valuation estimates? Another comment.' *Land Economics* 75(3), 478–82

Krawczyk, Michal, and Fabrice Le Lec (2015) 'Can we neutralize social preference in experimental games?' *Journal of Economic Behavior & Organization* 117, 340–55

Kuhn, Michael A., Peter Kuhn and Marie-Claire Villeval (2014) 'Self control and intertemporal choice: evidence from glucose and depletion interventions.' *CESifo working paper*

Lahey, Joanna N., and Douglas Oxley (2016) 'The power of eye tracking in economics experiments.' *American Economic Review* 106(5), 309–13

Laibson, David I. (1997) 'Golden eggs and hyperbolic discounting.' *Quarterly Journal of Economics* 112(2), 443–77

Langer, Thomas, and Craig R. Fox (2005) 'Biases in allocation under risk and uncertainty: partition dependence, unit dependence, and procedure dependence.' *University of Muenster, University of California at Los Angeles working paper*

Larrick, Richard P., and Jack B. Soll (2008) 'The mpg illusion.' *Science* 320(5883), 1593–94

Lazear, Edward P. (2000) 'Performance pay and productivity.' *American Economic Review* 90(5), 1346–61

Lazear, Edward P., Ulrike Malmendier and Roberto Weber (2011) 'Sorting in experiments with application to social preferences.' *American Economic Journal: Microeconomics* 4(1), 136–63

Lazear, Edward P., and Sherwin Rosen (1981) 'Rank-order tournaments as optimum labor contracts.' *Journal of Political Economy* 89(5), 841–64

Leamer, Edward E. (1983) 'Let's take the con out of econometrics.' *American Economic Review* 73(1), 31–43

LeBoeuf, Robyn A. (2006) 'Discount rates for time versus dates: the sensitivity of discounting to time-interval description.' *Journal of Marketing Research* 43(1), 59–72

Ledyard, John O. (1995) 'Public goods: a survey of experimental research.' In *Handbook of Experimental Economics,* ed. John H. Kagel and Alvin E. Roth (Princeton, NJ: Princeton University Press) pp. 111–94

Lee, Jinkwon (2008) 'The effect of the background risk in a simple chance improving decision model.' *Journal of Risk and Uncertainty* 36(1), 19–41

Lefebvre, Mathieu, Ferdinand M. Vieider and Marie-Claire Villeval (2010) 'Incentive effects on risk attitude in small probability prospects.' *Economics Letters* 109(2), 115–20

Lejuez, Carl W., Jennifer P. Read, Christopher W. Kahler, Jerry B. Richards, Susan E. Ramsey, Gregory L. Stuart, David R. Strong and Richard A. Brown (2002) 'Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (BART).' *Journal of Experimental Psychology: Applied* 8(2), 75–84

Levav, Jonathan, Mark Heitmann, Andreas Herrmann and Sheena S. Iyengar (2010) 'Order in product customization decisions: evidence from field experiments.' *Journal of Political Economy* 118(2), 274–99

Levine, David K., and Jie Zheng (2015) 'The relationship between economic theory and experiments.' In *Handbook of Experimental Economic Methodology,* ed. Guillaume Fréchette and Andrew Schotter (New York: Oxford University Press), 43–57

Levitt, Steven D. and John A. List (2007) 'What do laboratory experiments measuring social preferences reveal about the real world?' *Journal of Economic Perspectives* 21(2), 153–74

—— (2011) 'Was there really a Hawthorne effect at the Hawthorne plant? An analysis of the original illumination experiments.' *American Economic Journal: Applied Economics* 3(1), 224–38

Lévy-Garboua, Louis, Hela Maafi, David Masclet and Antoine Terracol (2012) 'Risk aversion and framing effects.' *Experimental Economics* 15(1), 128–44

Lezzi, Emanuela, Piers Fleming and Daniel J. Zizzo (2015) 'Does it matter which effort task you use? A comparison of four effort tasks when agents compete for a prize.' *University of East Anglia working paper*

L'Haridon, Olivier, and Ferdinand M. Vieider (2015) 'All over the map: heterogeneity of risk preferences across individuals, contexts, and countries.' *Working paper*

Li, Chen, Zhihua Li and Peter P. Wakker (2014) 'If nudge cannot be applied: a litmus test of the readers' resistance on paternalism.' *Theory and Decision* 76(3), 297–315

Liberati, Alessandro, Douglas G. Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C. Gotzsche, John P.A. Ioannidis, Mike Clarke, P.J. Devereaux, Jos Kleijnen and David Moher (2009) 'The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration.' *PLoS Medicine* 6(7), e1000100

Lichtenstein, Sarah, and Paul Slovic (1971) 'Reversals of preference between bids and choices in gambling decisions.' *Journal of Experimental Psychology* 89(1), 46–55

—— (1973) 'Response-induced reversals of preference in gambling: an extended replication in Las Vegas.' *Journal of Experimental Psychology* 101(1), 16–20

List, John A. (2001) 'Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards.' *American Economic Review* 91(5), 1498–1507

—— (2003) 'Does market experience eliminate market anomalies?' *Quarterly Journal of Economics* 118(1), 41–72

—— (2004) 'Neoclassical theory versus prospect theory: evidence from the marketplace.' *Econometrica* 72(2), 615–25

List, John A., Charles D. Bailey, Patricia J. Euzent and Thomas L. Martin (2001) 'Academic economists behaving badly? A survey on three areas of unethical behavior.' *Economic Inquiry* 39(1), 162–70

List, John A., and Craig A. Gallet (2001) 'What experimental protocol influence disparities between actual and hypothetical stated values?' *Environmental & Resource Economics* 20(3), 241–54

Loewenstein, George (1999) 'Experimental economics from the vantage-point of behavioural economics.' *Economic Journal* 109(453), F25–F34

Loewenstein, George, and Drazen Prelec (1992) 'Anomalies in intertemporal choice: evidence and an interpretation.' *Quarterly Journal of Economics* 107(2), 573–97

Loomes, Graham, Peter G. Moffatt and Robert Sugden (2002) 'A microeconometric test of alternative stochastic theories of risky choice.' *Journal of Risk and Uncertainty* 24(2), 103–30

Loomes, Graham, and Robert Sugden (1995) 'Incorporating a stochastic element into decision theories.' *European Economic Review* 39(3), 641–8

Loomis, John, Armando Gonzalez-Caban and Robin Gregory (1994) 'Do reminders of substitutes and budget constraints influence contingent valuation estimates?' *Land Economics* 70(4), 499–506

Luce, Duncan R. (1959) *Individual Choice Behavior* (New York: Wiley)

—— (1991) 'Rank- and sign-dependent linear utility models for binary gambles.' *Journal of Economic Theory* 53(1), 75–100

—— (2014) *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches* (New York: Psychology Press)

Lusk, Jayson L. (2003) 'Effects of cheap talk on consumer willingness-to-pay for golden rice.' *American Journal of Agricultural Economics* 85(4), 840–56

Lynch, John G., Jr, and Dan Ariely (2000) 'Wine online: search costs affect competition on price, quality, and distribution.' *Marketing Science* 19(1), 83–103

McCord, Mark, and Richard Neufville (1985) 'Assessment response surface: investigating utility dependence on probability.' *Theory and Decision* 18(3), 263–85

McFadden, Daniel (1999) 'Rationality for economists?' *Journal of Risk and Uncertainty* 19(1), 73–105

McKelvey, Richard D., and Talbot Page (1990) 'Public and private information: an experimental study of information pooling.' *Econometrica* 58(6), 1321–39

McKelvey, Richard D., and Thomas R. Palfrey (1992) 'An experimental study of the centipede game.' *Econometrica* 60(4), 803–36

—— (1995) 'Quantal response equilibria for normal form games.' *Games and Economic Behavior* 10, 6–38

—— (1998) 'Quantal response equilibria for extensive form games.' *Experimental Economics* 1(1), 9–41

McQuillin, Ben, and Robert Sugden (2012) 'Reconciling normative and behavioural economics: the problems to be solved.' *Social Choice and Welfare* 38(4), 553–67

Maccheroni, Fabio, Massimo Marinacci and Aldo Rustichini (2006) 'Ambiguity aversion, robustness, and the variational representation of preferences.' *Econometrica* 74(6), 1447–98

Machina, Mark J. (1982) '"expected utility" analysis without the independence axiom.' *Econometrica* 50(2), 277–323

—— (1987) 'Choice under uncertainty: problems solved and unsolved.' *Journal of Economic Perspectives* 1(1), 121–54

Machina, Mark J., and David Schmeidler (1992) 'A more robust definition of subjective probability.' *Econometrica* 60(4), 745–80

—— (1995) 'Bayes without Bernoulli: Simple conditions for probabilistically sophisticated choice.' *Journal of Economic Theory* 67(1), 106–28

Madrian, Brigitte C., and Dennis F. Shea (2001) 'The power of suggestion: inertia in 401(k) participation and savings behavior.' *Quarterly Journal of Economics* 116(4), 1149–87

Madsen, Dag O., and Tonny Stenheim (2015) 'Experimental methods in economics and psychology: a comparison.' *Procedia: Social and Behavioral Sciences* 187, 113–17

Malcomson, James M. (1984) 'Work incentives, hierarchy, and internal labor markets.' *Journal of Political Economy* 92(3), 486–507.

Maniadis, Zacharias, Fabio Tufano and John A. List (2014) 'One swallow doesn't make a summer: new evidence on anchoring effects.' *American Economic Review* 104(1), 277–90

Manski, Charles F. (1999) *Identification Problems in the Social Sciences* (Cambridge, MA: Harvard University Press)

—— (2002) 'Identification of decision rules in experiments on simple games of proposal and response.' *European Economic Review* 46(4–5), 880–91

—— (2004) 'Measuring expectations.' *Econometrica* 72(5), 1329–76

—— (2006) 'Interpreting the predictions of prediction markets.' *Economics Letters* 91(3), 425–9

Manzini, Paola, and Marco Mariotti (2014) 'A case of framing effects: the elicitation of time preferences.' *University of St Andrews working paper*

Marshall, Alfred (1890) *Principles of Economics* (London: Macmillan and Company)

Martin, Jolie M., and Michael I. Norton (2009) 'Shaping online consumer choice by partitioning the Web.' *Psychology & Marketing* 26(10), 908–26

Masclet, David, Charles N. Noussair, Steven Tucker and Marie-Claire Villeval (2003) 'Monetary and nonmonetary punishment in the voluntary contributions mechanism.' *American Economic Review* 93(1), 366–80

Massoni, Sébastien, Thibault Gajdos and Jean-Christophe Vergnaud (2014) 'Confidence measurement in the light of signal detection theory.' *Frontiers in Psychology* 5(1455), 1–13

Mayo, Elton (1949) *Hawthorne and the Western Electric Company: The Social Problems of an Industrial Civilization* (New York: Routledge)

Meidinger, Claude, Jean-Louis Rullière and Marie-Claire Villeval (2003) 'Does team-based compensation give rise to problems when agents vary in their ability?' *Experimental Economics* 6(3), 253–72

Mill, John Stuart (1836) 'On the definition of political economy and the method of investigation proper to it.' In *Collected works of John Stuart Mill*, vol. 4 (Toronto: University of Toronto Press) pp. 120–64

Mischel, Walter, Yuichi Shoda and Monica L. Rodriguez (1989) 'Delay of gratification in children.' *Science* 244(4907), 933–8

Mitchell, Gregory (2005) 'Libertarian paternalism is an oxymoron.' *Northwestern University Law Review* 99(3), 1245–77

Miyamoto, John M. (1988) 'Generic utility theory: measurement foundations and applications in multiattribute utility theory.' *Journal of Mathematical Psychology* 32(4), 357–404

Moffatt, Peter G. (2015) *Experimetrics: Econometrics for Experimental Economics* (London: Macmillan Education UK)

Moffatt, Peter G., and Simon Peters (2001) 'Testing for the presence of a tremble in economic experiments.' *Experimental Economics* 4(3), 221–8

Montmarquette, Claude, Jean-Louis Rullière, Marie-Claire Villeval and Romain Zeiliger (2004) 'Redesigning teams and incentives in a merger: an experiment with managers and students.' *Management Science* 50(10), 1379–89

Morrison, Gwendolyn (2000) 'The endowment effect and expected utility.' *Scottish Journal of Political Economy* 47(2), 183–97

Morrison, Mark Daniel, and Thomas C. Brown (2009) 'Testing the effectiveness of certainty scales, cheap talk, and dissonance-minimization in reducing hypothetical bias in contingent valuation studies.' *Environmental & Resource Economics* 44(3), 307–26

Munnell, Alicia H. (2003) 'A non-libertarian paternalist's reaction to "Libertarian paternalism is not an oxymoron".' *Federal Reserve Bank of Boston 48th Annual Conference*

Murnighan, Keith J., Alvin E. Roth and Françoise Schoumaker (1988) 'Risk aversion in bargaining: an experimental study.' *Journal of Risk and Uncertainty* 1(1), 101–24

Murphy, James J., Thomas H. Stevens and Darryl Weatherhead (2005) 'Is cheap talk effective at eliminating hypothetical bias in a provision point mechanism?' *Environmental & Resource Economics* 30(3), 327–43

Murphy, James J., Thomas H. Stevens and Lava Yadav (2010) 'A comparison of induced value and home-grown value experiments to test for hypothetical bias in contingent valuation.' *Environmental & Resource Economics* 47(1), 111–23

Nagel, Rosemarie (1995) 'Unraveling in guessing games: an experimental study.' *American Economic Review* 85(5), 1313–26

Nalbantian, Haig R., and Andrew Schotter (1995) 'Matching and efficiency in the baseball free-agent system: an experimental examination.' *Journal of Labor Economics* 13(1), 1–31

Neill, Helen R., Ronald G. Cummings, Philip T. Ganderton, Glenn W. Harrison and Thomas McGuckin (1994) 'Hypothetical surveys and real economic commitments.' *Land Economics* 70(2), 145–54

Neri, Claudia (2015) 'Eliciting beliefs in continuous-choice games: a double auction experiment.' *Experimental Economics* 18(4), 1–40

Normann, Hans-Theo, Till Requate and Israel Waichman (2014) 'Do short-term laboratory experiments provide valid descriptions of long-term economic interactions? A study of cournot markets.' *Experimental Economics* 17(3), 371–90

Normann, Hans-Theo, and Roberto Ricciuti (2009) 'Laboratory experiments for economic policy making.' *Journal of Economic Surveys* 23(3), 407–32

Normann, Hans-Theo, and Brian Wallace (2012) 'The impact of the termination rule in cooperation experiments.' *International Journal of Game Theory* 41(3), 707–18

Noussair, Charles (2011) 'Trends in academic publishing in experimental economics.' *Journal of Economic Survey, online conference*

Noussair, Charles N., Stéphane Robin and Bernard Ruffieux (2004) 'Revealing consumers' willingness-to-pay: a comparison of the BDM mechanism and the Vickrey auction.' *Journal of Economic Psychology* 25(6), 725–41

Noussair, Charles N., and Daan van Soest (2014) 'Experimental approaches to resource and environmental economics.' *Annual Reviews of Resource Economics* 6, 319–37

Nyarko, Yaw, and Andrew Schotter (2002) 'An experimental study of belief learning using elicited beliefs.' *Econometrica* 70(3), 971–1005

O'Donoghue, Ted, and Matthew Rabin (1998) 'Procrastination in preparing for retirement.' In *Behavioral dimensions of retirement economics,* ed. Henry Aaron (Washington, DC Brookings Institution Press), 125–56

—— (2001) 'Choice and procrastination.' *Quarterly Journal of Economics* 116(1), 121–60

Offerman, Theo, and Asa B. Palley (2016) 'Lossed in translation: an off-the-shelf method to recover probabilistic beliefs from loss-averse agents.' *Experimental Economics* 19(1), 1–30

Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen and Peter P. Wakker (2009) 'A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes.' *Review of Economic Studies* 76(4), 1461–89

Olken, Benjamin A. (2015) 'Promises and perils of pre-analysis plans.' *Journal of Economic Perspectives* 29(3), 61–80

Onay, Selçuk, and Ayse Öncüler (2007) 'Intertemporal choice under timing risk: an experimental approach.' *Journal of Risk and Uncertainty* 34(2), 99–121

Öncüler, Ayse (2000) 'Intertemporal choice under uncertainty: a behavioral perspective.' *Unpublished dissertation*

Open Science Collaboration (2015) 'Estimating the reproducibility of psychological science.' *Science* 349(6251), aac4716

Ortmann, Andreas (2010) 'The way in which an experiment is conducted is unbelievably important: on the experimentation practices of economists and psychologists.' *University of New South Wales working paper*

Ortmann, Andreas, and Ralph Hertwig (2002) 'The costs of deception: evidence from psychology.' *Experimental Economics* 5(2), 111–31

Osborne, Martin J., and Ariel Rubinstein (1994) *A Course in Game Theory* (Cambridge, MA: MIT Press)

Oxoby, Robert J., and John M. Spraggon (2008) 'Mine and yours: property rights in dictator games.' *Journal of Economic Behavior & Organization* 65(3–4), 703–13

Pais, Joana, and Ágnes Pintér (2008) 'School choice and information: an experimental study on matching mechanisms.' *Games and Economic Behavior* 64(1), 303–28

Pais, Joana, Agnes Pintér and Robert F. Veszteg (2012) 'Decentralized matching markets: a laboratory experiment.' *University of Lisbon, ISEG working paper*

Palacios-Huerta, Ignacio, and Oscar Volij (2008) 'Experientia Docet: professionals play minimax in laboratory experiments.' *Econometrica* 76(1), 71–115

Paolacci, Gabriele, and Jesse Chandler (2014) 'Inside the Turk: understanding the mechanical Turk as a participant pool.' *Current Directions in Psychological Science* 23(3), 184–8

Paolacci, Gabriele, Jesse Chandler and Panagiotis G. Ipeirotis (2010) 'Running experiments on Amazon mechanical Turk.' *Judgment and Decision Making* 5(5), 411–19

Parkhurst, Gregory M., and Jason F. Shogren (2005) 'Does complexity reduce coordination?' *Applied Economics Letters* 12(7), 447–52

Paserman, M. Daniele (2008) 'Job search and hyperbolic discounting: structural estimation and policy evaluation.' *Economic Journal* 118(531), 1418–52

Peters, Ellen, Nathan F. Dieckmann, Daniel Västfjäll, C.K. Mertz, Paul Slovic and Judith H. Hibbard (2009) 'Bringing meaning to numbers: the impact of evaluative categories on decisions.' *Journal of Experimental Psychology: Applied* 15(3), 213–27

Phelps, Edmund S., and Robert A. Pollak (1968) 'On second-best national saving and game-equilibrium growth.' *Review of Economic Studies* 35(2), 185–99

Piovesan, Marco, and Erik Wengstrom (2009) 'Fast or fair? A study of response times.' *Economics Letters* 105(2), 193–6

Plott, Charles R. (1982) 'Industrial organization theory and experimental economics.' *Journal of Economic Literature* 20(4), 1485–1527

—— (1989) 'An updated review of industrial organization: applications of experimental methods.' In *Handbook of Industrial Organization,* ed. Richard Schmalensee and Robert Willig, vol. 2 (Amsterdam: North Holland) pp. 1109–76

—— (1991) 'Will economics become an experimental science?' *Southern Economic Journal* 57(4), 901–19

Plott, Charles R., and Vernon L. Smith (2008) *Handbook of Experimental Economics Results,* vol. 1 (Amsterdam: Elsevier)

Plott, Charles R., and Kathryn Zeiler (2005) 'The willingness to pay–willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations.' *American Economic Review* 85(3), 530–45

—— (2007) 'Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory?' *American Economic Review* 97(4), 1449–66

——(2011) 'The willingness to pay–willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: reply.' *American Economic Review* 101(2), 1012–28

Poe, Gregory L., Jeremy E. Clark, Daniel Rondeau and William D. Schulze (2002) 'Provision point mechanisms and field validity tests of contingent valuation.' *Environmental & Resource Economics* 23(1), 105–31

Prelec, Drazen (1998) 'The probability weighting function.' *Econometrica* 66(3), 497–528

——(2004) 'A Bayesian truth serum for subjective data.' *Science* 306(5695), 462–6

Qiu, Jianying, and Eva-Maria Steiger (2011) 'Understanding the two components of risk attitudes: an experimental analysis.' *Management Science* 57(1), 193–9

Quiggin, John (1982) 'A theory of anticipated utility.' *Journal of Economic Behavior & Organization* 3(4), 323–43

Rabin, Matthew (1993) 'Incorporating fairness into game theory and economics.' *American Economic Review* 83(5), 1281–1302

——(1998) 'Psychology and economics.' *Journal of Economic Literature* 36(1), 11–46

——(2000) 'Risk aversion and expected-utility theory: a calibration theorem.' *Econometrica* 68(5), 1281–92

——(2013) 'Incorporating limited rationality into economics.' *Journal of Economic Literature* 51(2), 528–43

Raiffa, Howard (1968) *Decision Analysis: Introductory Lectures on Choices Under Uncertainty* (Reading, MA Addison-Wesley)

Ramsey, Frank P. (1931) 'Truth and probability.' In *The Foundations of Mathematics and other Logical Essays,* ed. Richard B. Braithwaite (London: Routledge and Kegan Paul), 156–98

Rand, David G. (2016) 'Cooperation, fast and slow.' *Psychological Science* 27(9), 1192–1206

Rand, David G., Joshua D. Green and Martin A. Nowak (2012) 'Spontaneous giving and calculated greed.' *Nature* 489(7416), 427–30

Rapoport, Amnon, William E. Stein, James E. Parco and Thomas E. Nicholas (2003) 'Equilibrium play and adaptive learning in a three-person centipede game.' *Games and Economic Behavior* 43(2), 239–65

Raven, John (2008) 'General introduction and overview. The Raven progressive matrices tests: their theoretical basis and measurement model.' In *Uses and abuses of intelligence,* ed. John and Jean Raven (Edinburgh: Competency Motivation Project), 17–68

Razali, Nornadiah Mohd, and Yap Bee Wah (2011) 'Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests.' *Journal of Statistical Modeling and Analytics* 2(1), 21–33

Read, Daniel, Shane Frederick, Burcu Orsel and Juwaria Rahman (2005) 'Four score and seven years from now: The date/delay effect in temporal discounting.' *Management Science* 51(9), 1326–35

Read, Daniel, and Peter H. Roelofsma (2003) 'Subadditive versus hyperbolic discounting: a comparison of choice and matching.' *Organizational Behavior and Human Decision Processes* 91(2), 140–53

Reny, Philip J. (1993) 'Common belief and the theory of games with perfect information.' *Journal of Economic Theory* 59(2), 257–74

Requate, Till, and Israel Waichman (2011) ' "a profit table or a profit calculator?" A note on the design of cournot oligopoly experiments.' *Experimental Economics* 14(1), 36–46

Rey-Biel, Pedro (2009) 'Equilibrium play and best response to (stated) beliefs in normal form games.' *Games and Economic Behavior* 65(2), 572–85

Rhode, Paul W., and Koleman S. Strumpf (2004) 'Historical presidential betting markets.' *Journal of Economic Perspectives* 18(2), 127–41

Roby, Thornton B. (1964) 'Belief states: a preliminary empirical study.' *DTIC working paper*

Rohde, Kirsten I.M. (2010) 'The hyperbolic factor: a measure of time inconsistency.' *Journal of Risk and Uncertainty* 41(2), 125–40

Rosenthal, Robert W. (1981) 'Games of perfect information, predatory pricing and the chain-store paradox.' *Journal of Economic Theory* 25(1), 92–100

Rosenthal, Robert W., and Robin M. DiMatteo (2001) 'Meta-analysis: recent developments in quantitative methods for literature reviews.' *Annual Review of Psychology* 52(1), 59–82

Rosenzweig, Mark R., and Kenneth I. Wolpin (2000) 'Natural "natural experiments" in economics.' *Journal of Economic Literature* 38(4), 827–74

Roth, Alvin E. (1988) 'Laboratory experimentation in economics: a methodological overview.' *Economic Journal* 98(393), 974–1031

—— (1993) 'The early history of experimental economics.' *Journal of the History of Economic Thought* 15(2), 184–209

—— (1994) 'Lets keep the con out of experimental: a methodological note.' *Empirical Economics* 19(2), 279–89

Roth, Alvin E., and Michael W. Malouf (1979) 'Game-theoretic models and the role of information in bargaining.' *Psychological Review* 86(6), 574–94

Roth, Alvin E., and Marilda Sotomayor (1992) 'Two-sided matching.' In *Handbook of Game Theory with Economic Applications,* ed. Kenneth J. Arrow and Michael D. Intriligator, vol. 1 (Amsterdam: Elsevier), pp. 485–541

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey and Geoffrey Iverson (2009) 'Bayesian *t* tests for accepting and rejecting the null hypothesis.' *Psychonomic Bulletin & Review* 16(2), 225–37

Roux, Catherine, and Christian Thöni (2015) 'Do control questions influence behavior in experiments?' *Experimental Economics* 18(2), 185–94

Rubin, Donald B. (1974) 'Estimating causal effects of treatments in randomized and nonrandomized studies.' *Journal of Educational Psychology* 66(5), 688–701

Rubinstein, Ariel (2003) ' "Economics and psychology"? The case of hyperbolic discounting.' *International Economic Review* 44(4), 1207–16

—— (2006a) 'Comments on behavioral economics.' In *Advances in Economic Theory*, ed. R. Blundell, W. K. Newey and T. Persson (Cambridge: Cambridge University Press), 246–54

—— (2006b) 'Instinctive and cognitive reasoning: A study of response times.' *Working paper*

—— (2007) 'Instinctive and cognitive reasoning: a study of response times.' *Economic Journal* 117(523), 1243–59

—— (2012) 'Response time and decision making: A "free" experimental study.' *Working paper*

—— (2013) 'Response time and decision making: an experimental study.' *Judgment and Decision Making* 8(5), 540–51

Rutström, Elisabet E., and Nathaniel T. Wilcox (2009) 'Stated beliefs versus inferred beliefs: a methodological inquiry and experimental test.' *Games and Economic Behavior* 67(2), 616–32

Sachs, Lothar (2012) *Applied Statistics: A Handbook of Techniques* (New York: Springer Science & Business Media)

Salanié, François, and Nicolas Treich (2006) 'Over-savings and hyperbolic discounting.' *European Economic Review* 50(6), 1557–70

Salant, Yuval, and Ariel Rubinstein (2008) '(a, f): choice with frames.' *Review of Economic Studies* 75(4), 1287–96

Samuelson, Larry (2005) 'Economic theory and experimental economics.' *Journal of Economic Literature* 43(1), 65–107

Samuelson, Paul A. (1937) 'A note on measurement of utility.' *Review of Economic Studies* 4(2), 155–61

Samuelson, Paul A., and William Nordhaus (1985) *Economics*, 12th edn (New York: McGraw-Hill)

—— (1992) *Economics*, 14th edn (New York: McGraw-Hill)

Samuelson, William, and Richard J. Zeckhauser (1988) 'Status quo bias in decision making.' *Journal of Risk and Uncertainty* 1(1), 7–59

Sanfey, Alan G. (2007) 'Social decision-making: insights from game theory and neuroscience.' *Science* 318(5850), 598–602

Savage, Leonard J. (1954) *The Foundations of Statistics* (Cambridge: Cambridge University Press)

—— (1971) 'Elicitation of personal probabilities and expectations.' *Journal of the American Statistical Association* 66(336), 783–801

Sayman, Serdar, and Ayse Öncüler (2008) 'An investigation of time inconsistency.' *Management Science* 55(3), 470–82

Schelling, Thomas C. (1960) *The strategy of conflict* (Cambridge, MA: Harvard University Press)

Schlag, Karl H., and Joel J. van der Weele (2015) 'A method to elicit beliefs as most likely intervals.' *Judgment and Decision Making* 10(5), 456–468

Schlag, Karl H., James Tremewan and Joel J. van der Weele (2013) 'A penny for your thoughts: a survey of methods for eliciting beliefs.' *Experimental Economics* 1(3), 1–34

Schmeidler, David (1989) 'Subjective probability and expected utility without additivity.' *Econometrica* 57(3), 571–87

Schmuckler, Mark A. (2001) 'What is ecological validity? A dimensional analysis.' *Infancy* 2(4), 419–36

Schotter, Andrew (2006) 'Strong and wrong: the use of rational choice theory in experimental economics.' *Journal of Theoretical Politics* 18(4), 498–511

—— (2015) 'On the relationship between economic theory and experiments.' In *Handbook of Experimental Economic Methodology,* ed. Guillaume R. Fréchette and Andrew Schotter (New York: Oxford University Press), 58–85

Schotter, Andrew, and Isabel Trevino (2012) 'Is response time predictive of choice? An experimental study of threshold strategies.' *New York University working paper*

—— (2014) 'Belief elicitation in the laboratory.' *Annual Review of Economics* 6(1), 103–28

Schram, Arthur (2005) 'Artificiality: the tension between internal and external validity in economic experiments.' *Journal of Economic Methodology* 12(5), 225–37

Schulz, Jonathan F., Urs Fischbacher, Christian Thöni and Verena Utikal (2014) 'Affect and fairness: dictator games under cognitive load.' *Journal of Economic Psychology* 41, 77–87

Selten, R. (1973) 'A simple model of imperfect competition, where four are few and six are many.' *International Journal of Game Theory* 2(1), 141–201

Selten, Reinhard (1967) 'Die Strategiemethode zur Erforschung des eingeschrankt rationalen Verhaltens im Rahmen eines Oligopolexperiments.' In *Beiträge zur experimentellen Wirtschaftsforschung,* ed. Heinz Sauermann (Tübingen: J.C.B. Mohr), 136–68

—— (1998) 'Axiomatic characterization of the quadratic scoring rule.' *Experimental Economics* 1(1), 43–62

Selten, Reinhard, Abdolkarim Sadrieh and Klaus Abbink (1999) 'Money does not induce risk neutral behavior, but binary lotteries do even worse.' *Theory and Decision* 46(3), 213–52

Shaffer, Juliet Popper (1995) 'Multiple hypothesis testing.' *Annual Review of Psychology* 46(1), 561–84

Shalvi, Shaul, Ori Eldar and Yoella Bereby-Meyer (2012) 'Honesty requires time (and lack of justifications).' *Psychological Science* 23(10), 1264–70

Shapiro, Carl, and Joseph E. Stiglitz (1984) 'Equilibrium unemployment as a worker discipline device.' *American Economic Review* 74(3), 433–44

Shavit, Tal, Doron Sonsino and Uri Benzion (2001) 'A comparative study of lotteries: evaluation in class and on the web.' *Journal of Economic Psychology* 22(4), 483–91

Shearer, Bruce (2004) 'Piece rates, fixed wages and incentives: evidence from a field experiment.' *Review of Economic Studies* 71(2), 513–34

Shogren, Jason F. (2006) 'A rule of one.' *American Journal of Agricultural Economics* 88(5), 1147–59

Shu, Suzanne B. (2008) 'Future-biased search: the quest for the ideal.' *Journal of Behavioral Decision Making* 21(4), 352–77

Shu, Suzanne B., and Ayelet Gneezy (2010) 'Procrastination of enjoyable experiences.' *Journal of Marketing Research* 47(5), 933–44

Siakantaris, Nikos (2000) 'Experimental economics under the microscope.' *Cambridge Journal of Economics* 24(3), 267–81

Siegel, Sidney (1957) 'Nonparametric statistics.' *American Statistician* 11(3), 13–19

Siniscalchi, Marciano (2009) 'Vector expected utility and attitudes toward variation.' *Econometrica* 77(3), 801–55

Sitzia, Stefania, and Robert Sugden (2011) 'Implementing theoretical models in the laboratory, and what this can and cannot achieve.' *Journal of Economic Methodology* 18(4), 323–43

Slonim, Robert, Carmen Wang, Ellen Garbarino and Danielle Merrett (2013) 'Opting-in: participation bias in economic experiments.' *Journal of Economic Behavior & Organization* 90, 43–70

Slovic, Paul, Dale Griffin and Amos Tversky (1990) 'Compatibility effects in judgment and choice.' In *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, ed. R.M. Hogarth (Chicago: The University of Chicago Press), 5–27

Smith, Cedric A.B. (1961) 'Consistency in statistical inference and decision.' *Journal of the Royal Statistical Society. Series B (Methodological)* 23(1), 1–37

—— (1966) 'Correction: consistency in statistical inference and decision.' *Journal of the Royal Statistical Society. Series B (Methodological)* 28(1), 1–37

Smith, Vernon L. (1962) 'An experimental study of competitive market behavior.' *Journal of Political Economy* 70(2), 111–37

—— (1976) 'Experimental economics: induced value theory.' *American Economic Review* 66(2), 274–9

—— (1982) 'Microeconomic systems as an experimental science.' *American Economic Review* 72(5), 923–55

Smith, Vernon L., and James M. Walker (1993) 'Monetary rewards and decision cost in experimental economics.' *Economic Inquiry* 31(2), 245–61

Soll, Jack B., Ralph L. Keeney and Richard P. Larrick (2013) 'Consumer misunderstanding of credit card use, payments, and debt: causes and solutions.' *Journal of Public Policy & Marketing* 32(1), 66–81

Sonnemans, Joep, and Theo Offerman (2001) 'Is the quadratic scoring rule really incentive compatible?' *CREED working paper*

Sopher, Barry, and Gary Gigliotti (1993) 'Intransitive cycles: rational choice or random error? An answer based on estimation of error rates with experimental data.' *Theory and Decision* 35(3), 311–36

Spetzler, Carl S., and Carl-Axel Stael von Holstein (1975) 'Probability econding in decision analysis.' *Management Science* 22(3), 340–58

Stahl, Dale O., and Paul W. Wilson (1995) 'On players' models of other players: theory and experimental evidence.' *Games and Economic Behavior* 10(1), 218–54

Stanley, Tom D., Hristos Doucouliagos, Margaret Giles, Jost H. Heckemeyer, Robert J. Johnston, Patrice Laroche, Jon P. Nelson, Martin Paldam, Jacques Poot, Geoff Pugh, Randall S. Rosenberger and Katja Rost (2013) 'Meta-analysis of economics research reporting guidelines.' *Journal of Economic Surveys* 27(2), 390–4

Stapel, Diederik (2014) *Faking Science: A True Story of Academic Fraud* (Nicholas J. L. Brown (trans.), available at http://nick.brown.free.fr/stapel)

Starmer, Chris (2000) 'Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk.' *Journal of Economic Literature* 38(2), 332–82

Starmer, Chris, and Robert Sugden (1991) 'Does the random-lottery incentive system elicit true preferences? An experimental investigation.' *American Economic Review* 81(4), 971–8

Stigler, George J. (1965) *Essays in the History of Economics* (Chicago: The University of Chicago Press)

Stott, Henry (2006) 'Cumulative prospect theory's functional menagerie.' *Journal of Risk and Uncertainty* 32(2), 101–30

Stranlund, John K., James J. Murphy and John M. Spraggon (2014) 'Price controls and banking in emissions trading: an experimental evaluation.' *Journal of Environmental Economics and Management* 68(1), 71–86

Strotz, Robert Henry (1955) 'Myopia and inconsistency in dynamic utility maximization.' *Review of Economic Studies* 23(3), 165–80

Stroup, Donna F., Jesse A. Berlin, Sally C. Morton, Ingram Olkin, Williamson G. David, Drummond Rennie, David Moher, Betsy J. Becker, Theresa Ann Sipe and Stephen B. Thacker (2000) 'Meta-analysis of observational studies in epidemiology: a proposal for reporting.' *JAMA* 283(15), 2008–12

Sugden, Robert (2005) 'Experiments as exhibits and experiments as tests.' *Journal of Economic Methodology* 12(2), 291–302

Sunstein, Cass R., and Richard H. Thaler (2003) 'Libertarian paternalism is not an oxymoron.' *University of Chicago Law Review* 70(4), 1159–1202

Takahashi, Hiromasa, Junyi Shen and Kazuhito Ogawa (2016) 'An experimental examination of compensation schemes and level of effort in differentiated tasks.' *Journal of Behavioral and Experimental Economics* 61, 12–19

Takeuchi, Kan (2011) 'Non-parametric test of time consistency: present bias and future bias.' *Games and Economic Behavior* 71(2), 456–78

Tanaka, Tomomi, Colin F. Camerer, and Quang Nguyen (2010) 'Risk and time preferences: linking experimental and household survey data from Vietnam.' *American Economic Review* 100(1), 557–71

Taylor, Laura O., Michael McKee, Susan K. Laury and Ronald G. Cummings (2001) 'Induced-value tests of the referendum voting mechanism.' *Economics Letters* 71(1), 61–5

Thaler, Richard H. (1981) 'Some empirical evidence on dynamic inconsistency.' *Economics Letters* 8(3), 201–7

Thaler, Richard H., and Shlomo Benartzi (2004) 'Save more tomorrow: using behavioral economics to increase employee saving.' *Journal of Political Economy* 112(1), S164–S187

Thaler, Richard H., and Eric J. Johnson (1990) 'Gambling with the house money and trying to break even: the effects of prior outcomes on risky choice.' *Management Science* 36(6), 643–60

Thaler, Richard H., and Cass R. Sunstein (2003) 'Libertarian paternalism.' *American Economic Review* 93(2), 175–9

—— (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness* (New Haven, CT: Yale University Press)

Toda, Masanao (1963) 'Measurement of subjective probability distributions.' *DTIC working paper*

Torgler, Benno (2002) 'Speaking to theorists and searching for facts: tax morale and tax compliance in experiments.' *Journal of Economic Surveys* 16(5), 657–83

Train, Kenneth (2009) *Discrete Choice Methods with Simulation* (Cambridge: Cambridge University Press)

Trautmann, Stefan T., and Gijs van de Kuilen (2014) 'Belief elicitation: a horse race among truth serums.' *Economic Journal* 125(589), 2116–35

—— (2016) 'Ambiguity attitudes.' *The Wiley Blackwell Handbook of Judgment and Decision Making* 1, 89–116

Trautmann, Stefan T., Ferdinand M. Vieider and Peter P. Wakker (2011) 'Preference reversals for ambiguity aversion.' *Management Science* 57(7), 1320–33

Tukey, John W. (1960) 'A survey of sampling from contaminated distributions.' *Contributions to Probability and Statistics* 2, 448–85

Tversky, Amos, and Daniel Kahneman (1974) 'Judgment under uncertainty: heuristics and biases.' *Science* 185(4157), 1124–31

—— (1986) 'Rational choice and the framing of decisions.' *Journal of Business* 59(4), S251–S278

—— (1992) 'Advances in prospect theory: cumulative representation of uncertainty.' *Journal of Risk and Uncertainty* 5(4), 297–323

Tversky, Amos, Paul Slovic and Daniel Kahneman (1990) 'The causes of preference reversal.' *American Economic Review* 80(1), 204–17

Tversky, Amos, and Peter P. Wakker (1995) 'Risk attitudes and decision weights.' *Econometrica* 63(6), 1255–80

van Assen, Marcel, and Chris Snijders (2010) 'The effect of nonlinear utility on behaviour in repeated prisoner's dilemmas.' *Rationality and Society* 22(3), 301–32

van Huyck, John B., Raymond C. Battalio and Richard O. Beil (1990) 'Tacit coordination games, strategic uncertainty, and coordination failure.' *American Economic Review* 80(1), 234–48

Vieider, Ferdinand M., Mathieu Lefebvre, Ranoua Bouchouicha, Thorsten Chmura, Rustamdjan Hakimov, Michal Krawczyk and Peter Martinsson (2015) 'Common components of risk and uncertainty attitudes across contexts and domains: evidence from 30 countries.' *Journal of the European Economic Association* 13(3), 421–52

von Gaudecker, Hans-Martin, Arthur van Soest and Erik Wengström (2011) 'Heterogeneity in risky choice behaviour in a broad population.' *American Economic Review* 101(2), 664–94

Von Neumann, John, and Oskar Morgenstern (1944) *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press)

Vossler, Christian A., and Michael McKee (2006) 'Induced-value tests of contingent valuation elicitation mechanisms.' *Environmental & Resource Economics* 35(2), 137–68

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El ghormli and Nathaniel Rothman (2004) 'Assessing the probability that a positive report is false: an approach for molecular epidemiology studies.' *Journal of the National Cancer Institute* 96(6), 434–42

Wakker, Peter P. (2004) 'On the composition of risk preference and belief.' *Psychological Review* 111(1), 236–41

—— (2010) *Prospect Theory for Risk and Ambiguity* (Cambridge: Cambridge University Press)

Wakker, Peter P., and Daniel Deneffe (1996) 'Eliciting Von Neumann–Morgenstern utilities when probabilities are distorted or unknown.' *Management Science* 42(8), 1131–50

Walker, Joe (1987) 'Experimental economics in the classroom.' *Journal of Economic Education* 18(1), 51–7

Wallis, W. Allen, and Milton Friedman (1942) 'The empirical derivation of indifference functions.' In *Studies in Mathematical Economics and Econometrics in Memory of Henry Schultz,* ed. O. Lange, F. McIntyre and T.O. Yntema (Chicago: The University of Chicago Press) pp. 175–89

Weaver, Ray, and Drazen Prelec (2013) 'Creating truthtelling incentives with the Bayesian truth serum.' *Journal of Marketing Research* 50(3), 289–302

Weber, Elke U., and Patricia G. Lindemann (2011) 'From intuition to analysis: making decisions with our head, our heart, or by the book.' In *Intuition in Judgment and Decision Making,* ed. Henning Plessner, Cornelia Betsch and Tilmann Betsch (New York: Taylor and Francis) pp. 191–208

Whitehead, John C., and Glenn C. Blomquist (1995) 'Do reminders of substitutes and budget constraints influence contingent valuation estimates? Comment.' *Land Economics* 71(4), 541–3

—— (1999) 'Do reminders of substitutes and budget constraints influence contingent valuation estimates? Reply to another comment.' *Land Economics* 75(3), 483–4

Wilcox, Nathaniel T. (2008) 'Stochastic models for binary discrete choice under risk: a critical primer and econometric comparison.' In *Research in Experimental Economics,* ed. James C. Cox and Glenn W. Harrison, vol. 12 (Bingley: Emerald Group) pp. 197–292

Winkler, Robert L., and Allan H. Murphy (1970) 'Nonlinear utility and the probability score.' *Journal of Applied Meteorology* 9(1), 143–8

Wolfers, Justin, and Eric Zitzewitz (2006) 'Interpreting prediction market prices as probabilities.' *NBER working paper*

Wooldridge, Jeffrey M. (2002) *Econometric Analysis of Cross-section and Panel Data* (Cambridge, MA: MIT Press)

Wu, George, and Richard Gonzalez (1996) 'Curvature of the probability weighting function.' *Management Science* 42(12), 1676–90

Yates, Frank J., and Lisa G. Zukowski (1976) 'Characterization of ambiguity in decision making.' *Behavioral Science* 21(1), 19–25

Zank, Horst (2010) 'On probabilities and loss aversion.' *Theory and Decision* 68(3), 243–61

Zimbardo, Philip (2007) *The Lucifer Effect: Understanding How Good People Turn Evil* (New York: Random House)

Zimmermann, Christian (2015) 'On the need for a replication journal.' *Federal Reserve Bank of St Louis working paper*

Zizzo, Daniel J. (2010) 'Experimenter demand effects in economic experiments.' *Experimental Economics* 13(1), 75–98

Index

Index of Authors