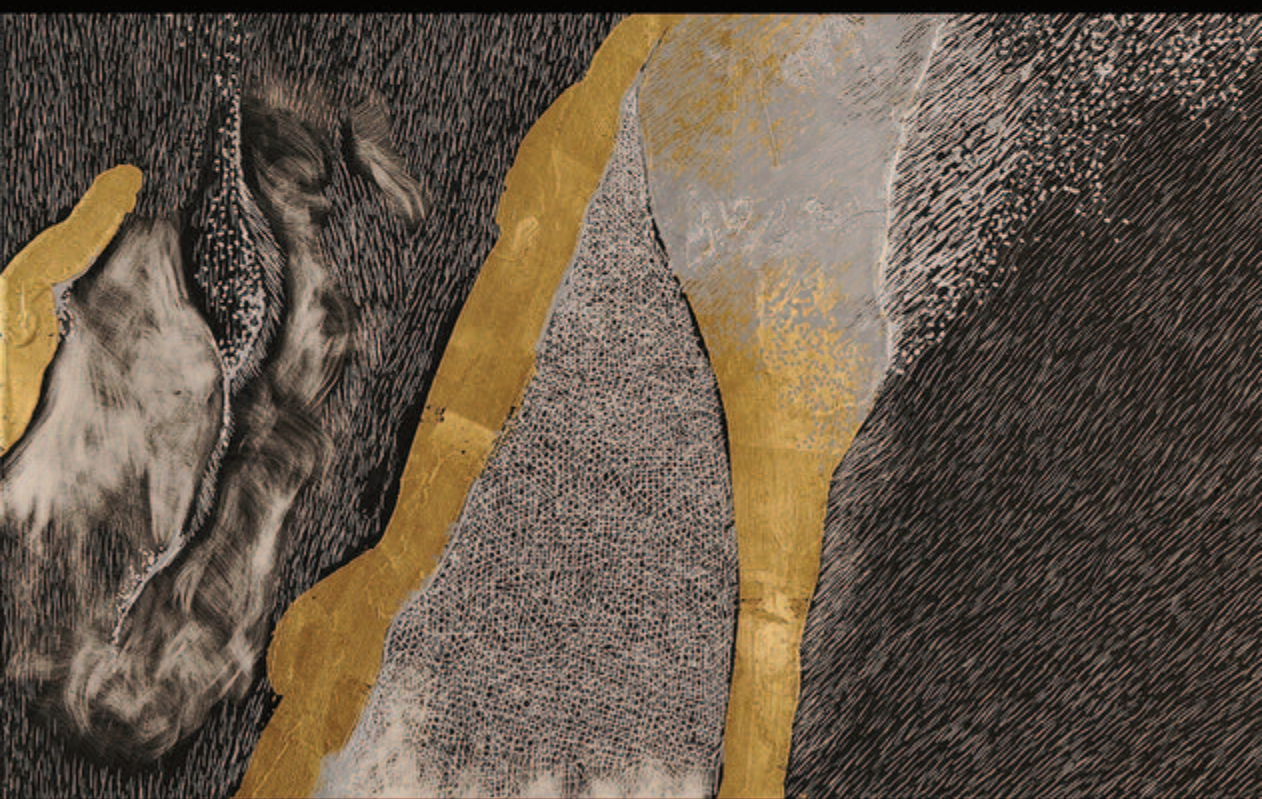
Handbook of Experimental Economic Methodology

Edited by

Guillaume R. Fréchette and Andrew Schotter



THE HANDBOOKS IN ECONOMIC METHODOLOGIES SERIES

In collaboration with the Center for Experimental Social Science, New York University

HANDBOOK OF EXPERIMENTAL ECONOMIC METHODOLOGY

The Handbooks in Economic Methodologies Series

The Foundations of Positive and Normative Economics: A Handbook
Edited by Andrew Caplin and Andrew Schotter

Handbook of Experimental Economic Methodology
Edited by Guillaume R. Fréchette and Andrew Schotter

In collaboration
with the Center for Experimental Social Science, New York University

HANDBOOK OF EXPERIMENTAL ECONOMIC METHODOLOGY

Edited by
GUILLAUME R. FRÉCHETTE
and
ANDREW SCHOTTER

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
Handbook of experimental economic methodology / edited by Guillaume R. Fréchette
and Andrew Schotter.
p. cm. — (The handbooks in economic methodologies series)
Includes bibliographical references and index.
ISBN 978–0–19–532832–5 (alk. paper)
1. Experimental economics. 2. Economics—Methodology. I. Fréchette, Guillaume R. II. Schotter, A.
HB131.H3547 2015
330.072—dc23
2014025000

1 3 5 7 9 8 6 4 2

Printed in the United States of America on acid-free paper

Contents

Contributors vii
Acknowledgments ix
Series Introduction xi

Introduction 1

PART I. INTRODUCTION

1. Is Experimental Economics Living Up to Its Promise?—Alvin E. Roth 13

PART II. ECONOMIC THEORY AND EXPERIMENTAL ECONOMICS

2. The Relationship Between Economic Theory and Experiments—David K. Levine and Jie Zheng 43
3. On the Relationship Between Economic Theory and Experiments—Andrew Schotter 58
4. Enhanced Choice Experiments—Andrew Caplin and Mark Dean 86
5. Intelligent Design: The Relationship Between Economic Theory and Experiments: Treatment-driven Experiments—Muriel Niederle 104
6. The Interplay Between Theory and Experiments—Leeat Yariv 132
7. Maxims for Experimenters—Martin Dufwenberg 141
8. What is an Economic Theory That Can Inform Experiments?—Uri Gneezy and Pedro Rey-Biel 145

PART III. PSYCHOLOGY AND ECONOMICS: A COMPARISON OF METHODS

9. The 1-800 Critique, Counterexamples, and the Future of Behavioral Economics—Ido Erev and Ben Greiner 151
10. A General Model for Experimental Inquiry in Economics and Social Psychology—J. Keith Murnighan 166
11. Psychology and Economics: Areas of Convergence and Difference—Tom R. Tyler and David M. Amodio 181

Shorter Papers and Comments

12. The Hammer and the Screwdriver—Gary Charness 197
13. Discussion of "Psychology and Economics: Areas of Convergence and Difference"—Theo Offerman 200

PART IV. THE LABORATORY AND THE FIELD

Reprint: What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?—Steven D. Levitt and John A. List 207
14. The Promise and Success of Lab–Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List—Colin F. Camerer 249
15. Theory, Experimental Design, and Econometrics Are Complementary (And So Are Lab and Field Experiments)—Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström 296
16. Laboratory Experiments: The Lab in Relationship to Field Experiments, Field Data, and Economic Theory—John H. Kagel 339
17. Laboratory Experiments: Professionals Versus Students—Guillaume R. Fréchette 360

Shorter Papers and Comments

18. The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects—Judd B. Kessler and Lise Vesterlund 391
19. The Lab and the Field: Empirical and Experimental Economics—David Reiley 407
20. On the Generalizability of Experimental Results in Economics—Omar Al-Ubaydli and John A. List 420

Index 463

Contributors

Omar Al-Ubaydli: Bahrain Center for Strategic, International, and Energy Studies, and George Mason University

David M. Amodio: New York University

Colin F. Camerer: California Institute of Technology

Andrew Caplin: New York University

Gary Charness: University of California, Santa Barbara

Mark Dean: Brown University

Martin Dufwenberg: Bocconi University, University of Arizona, and University of Gothenburg

Ido Erev: Technion—Israel Institute of Technology

Guillaume R. Fréchette: New York University

Uri Gneezy: Rady School of Management at University of California San Diego and the Amsterdam School of Economics

Ben Greiner: University of New South Wales

Glenn W. Harrison: Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University

John H. Kagel: Ohio State University

Judd B. Kessler: The Wharton School, University of Pennsylvania

Morten I. Lau: Copenhagen Business School

David K. Levine: European University Institute and Washington University in St. Louis

Steven D. Levitt: University of Chicago

John A. List: Department of Economics, University of Chicago, and National Bureau of Economic Research

J. Keith Murnighan: Kellogg School of Management, Northwestern University

Muriel Niederle: Stanford University

Theo Offerman: University of Amsterdam

David Reiley: Google, Inc.

Pedro Rey-Biel: Universitat Autònoma de Barcelona and Barcelona GSE

Alvin E. Roth: Harvard University, Harvard Business School, and Stanford University

E. Elisabet Rutström: Dean's Behavioral Economics Laboratory, Robinson College of Business, and Andrew Young School of Policy Studies, Georgia State University

Andrew Schotter: New York University

Tom R. Tyler: Yale Law School, Yale University

Lise Vesterlund: University of Pittsburgh

Leeat Yariv: California Institute of Technology

Jie Zheng: School of Economics and Management, Tsinghua University

Acknowledgments

There are many people and organizations we would like to thank who have been instrumental in bringing this book to completion. First and foremost, we would like to thank Eric Wanner and the Russell Sage Foundation for their financial support, and we would like to acknowledge the support of the Center for Experimental Social Science at New York University. In addition, we owe a great debt to Daniel Martin, who laboriously handled the production and editing of the manuscript and Nic Kozeniauskas who shepherded to book through its final stages. We also would like to thank Terry Vaughn of Oxford University Press for encouraging us to produce this volume; and we are grateful to Scott Paris, who took over for Terry. Last, but certainly not least, we would like to thank our contributing authors not only for their lively chapters and presentations but also for the patience they exhibited in waiting for this volume to finally be published.

Series Introduction

Durable contributions to economic thought that focus purely on methodology are few and far between. Absent substantive progress, methodological reflections are widely seen as fruitless, while the methodological importance of successful research is seen as self-evident. Yet as the chapters written for this book attest, there are increasing signs of methodological ferment within the profession. We see this as long overdue, and for that reason we are editing a series of handbooks in a deliberate effort at raising methodological awareness. We believe that the next few decades will produce important advances in social scientific methodology. Improvements in our professional knowledge base and in the technology that is available to us have radically expanded potential avenues of research. By engaging in a collective discussion of the newly feasible options, we will improve our ability to exercise and explore the most fruitful among them. While we will surely start out on a wide variety of different research paths, a new consensus will form far more rapidly if we appreciate the motives of those forging these distinct routes.

While the direct goal of these handbooks is to engage with current researchers as they contemplate new research opportunities, we aim also to reach some who have yet to set out on their research journeys. Moreover, we hope that the books will be of interest to researchers in disciplines other than economics. We view the expansion of research technology as opening the door to the formation of research teams that cross traditional field boundaries. We are confident enough in the value of our intellectual traditions to believe that a well-thought-through and fully articulated "economic" methodology would serve as the best center point for such endeavors.

INTRODUCTION

Economists, like researchers in many other scientific fields, rarely stop to consider their methods. One view is that those who can do science, do it, whereas those who can't do it, talk about it. This is a volume filled with chapters written by some of the most accomplished scholars working at the intersection of experimental, behavioral, and theoretical economics talking about methodology.

We think that the time is right for such a discussion because experimental economics is no longer an emerging discipline but rather one that has matured to the point where it has become integrated into the thinking and curriculum of graduate training. However, along with growth and maturity comes a set of issues that any growing discipline has to confront. For example, as experimental work attempts to test theory, it raises the following question: What is the proper relationship between theory and experiments? As experimental results are used to inform policy, the question of their usefulness outside the lab is raised; and finally, as experimental economics tries to integrate ideas from other disciplines like psychology and neuroscience, the question of their proper place in our discipline arises.

This book is divided into four sections, three of which offer a discussion of the issues raised above. In addition, the volume starts off with a chapter by Al Roth, a veteran experimentalist and Nobel laureate, whose topic is whether experimental economics has lived up to its promise.

It is our hope that this volume will lead its readers to pause and think about the methods they are using in their everyday work and also about where the future of our discipline lies. Some of the chapters are contentious, which we feel is a healthy sign of a dynamic discipline, while others lay out a vision for how the authors think our discipline should be pursued. We feel that the sum total is a very exciting and illuminating collection of chapters on a topic at the core of experimental economics.

In each section of the book we have a set of chapters and a set of comments on those chapters. Our aim was to offer a place where ideas about methodology could be discussed, and we invited comments for that purpose. However, we gave our commenters a great deal of freedom not only to discuss the chapter they were invited to discuss but also to use that chapter as a jumping off point for their opinions. As you will see, our commenters took us up on both suggestions.

In this short introduction, we hope to give a precise overview of the main themes running through the discussion. Hence we by no means are attempting a summary of each chapter but rather a summary of what we think are the main areas of debate and how the chapters in the volume comment on them. In addition to the original chapters written exclusively for this volume, we are also reprinting the paper "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" by Steven Levitt and John A. List, a paper that has sparked a large debate about the relative merits of lab and field experiments. Since many of the chapters in the section on field and lab experiments comment on this paper, we felt it was appropriate to give readers easy access to the original Levitt–List paper so they can refer to it as they feel necessary.

Part I: Introduction

Our volume starts with the essay by Al Roth, who was asked to comment on the question of whether experimental economics is living up to its promise. To answer this question, Roth identifies three criteria: Has experimental economics lived up to its promise in informing theory, discovering new facts, and enlightening policy (what he calls "whispering in the ears of princes"). By and large the answer is yes and Roth provides a set of examples where, with the help of experiments, theory has gone through a process of "creative destruction" from theory testing, to exploring unpredicted regularities, to theory building and testing again. To illustrate this process, as Roth does, one might only need to follow the evolution of bargaining theory over its lifetime from the first efforts by Roth and collaborators to test the Nash bargaining theory, to the emergence of the Rubinstein bargaining theory, and finally to the theories of inequality aversion of Fehr and Schmidt and Bolton and Ocenfelds. Hence bargaining theory progressed by discovering new facts along the way—that is, that subgame perfect equilibrium was not predictive in the Ultimatum Game or that disadvantageous counter offers were robustly made in the shrinking pie game of Rubenstein. Finally, in terms of informing policy, the great advances in the theory and application of market design were clearly made with the aid of experiments which allowed experimental economists to whisper in the ear of princes.

Part II: Economic Theory and Experimental Economics

This part starts with two chapters that present rather different views of the relationship between theory and experiments, yet are united in their assessment of the need for theory to be able to predict outcomes in situations where no theory exists. Levine and Zheng take a relatively positive view of how successful economic theory has been when tested in the lab, stating that it has been more successful than it has been given credit for. When experimental data fail to support a theory, however, Levine and Zheng offer a number of examples where what has failed is not the theory but rather the interpretation given to it.

The second chapter, by Schotter, views theory in a different light. It suggests that economic theory is both "strong and wrong." It is strong in the sense that it is able to make precise predictions about behavior in many different situations, but is wrong in the sense that it is constructed to be logically and internally consistent with little attention given to whether the people it is tested on conform to the axioms of the theory. Furthermore, historically, the yardstick used to measure theory has been its elegance rather than its ability to predict behavior, and most theorists do not create their theories with an eye to having them tested as is true in Physics. Hence, as Schotter claims, we should not look to theory as a vehicle to explain the world of flesh and blood people but rather as a tool to help us structure our search for the truth. What is interesting are the robust systematic deviations from theory that need to be explained.

The chapter by Caplin and Dean asks us to broaden the type of data we generate and analyze in our experiments. While most theories make equilibrium predictions about choice, few offer a model of the process generating that choice. Clearly, however, if we are to fully understand how people make choices, we must be able to explain the process by which these choices are made. Caplin and Dean offer an example of how that can be done using a very clever design. The chapter offers us an important insight into how we can generate data that will inform us about the choice process and makes an important case for including nonstandard data, data that are not choice-based, into our repertoire. The benefit of including nonstandard data and the arguments against it was the focus of the first volume in this series entitled *The Foundations of Positive and Normative Economics.*

Niederle's chapter takes a more hands-on approach in the sense of rather than contemplating the relationship between theory and experiments, she offers an extremely useful blueprint for the proper testing of theory in the lab. She tries to give an answer to the question, What is a good experiment? by answering that a good experiment is one "…that allows testing for the main effect {of the theory} while controlling for other plausible alternatives. This helps ensure that the original hypothesis is reached for the right reasons and the initial theory is not wrongly confirmed." Niederle proposes that we test theory using direct tests which reveal

the properties of the theory in stark terms and, hence, reduce the need for heavy econometric inference. What is needed here, however, are intelligent experimental designs which provide these direct tests, and Neiderle provides us with many examples such designs.

Another way to look at the interaction between theory and experiments is to take the perspective of a sociologist of science and look at the collaborations of theorists and experimentalists, the schools they are attached to, and where they publish. If theory and experiments are merging, we should see it in the increased incidence of theorist–experimentalist collaborations and the increase in the number of theorists who add experiments to their repertoire. Leeat Yariv looks at these data. She finds that there is a substantial collaboration between theorists and experimentalists and a dramatic increase of theorists dabbling in experiments, especially among younger theorists, which is an encouraging sign.

Martin Dufwenberg offers advice to experimentalists on the proper way to go about creating an experiment and publishing it. He stresses the fact that often what is important about an experimental paper are the design and hypothesis being tested. The paper should be judged on the basis of the questions asked, and the design used to answer it, not on the basis of the results.

Finally, Uri Gneezy and Pedro Rey-Biel's starting point is the observation that while in economics a theory is considered to be a precise mathematical model of a phenomenon, in other social sciences quite different objects are considered to be theories. Hence, they suggest that it is not clear that experimentalists should limit themselves exclusively to testing only mathematical models. In addition, there is no need for economists to restrict themselves to only testing economic models or theories; we should consider theories from other fields and from empirically relevant environments, as well as use the laboratory as a "wind tunnel." However, they argue that experiments should be kept simple to better understand what participants are doing, and because of that, recovering parameters of a theoretical model from a laboratory experiment is rarely informative. Rather, the focus should be on treatment effects, which help understand the reasons behind a phenomenon and the direction of the effect.

Part III: Psychology and Economics: A Comparison of Methods

Because experimental economics is experimental, it shares many of the techniques and concerns of other experimental social sciences—in particular, those of psychology. Despite this commonality of method, there are still considerable differences in the way economists and psychologists go about their trades. This part of the book aims to explore these differences.

Erev and Greiner start Part III by contrasting the different approaches taken by these two disciplines. While economists create models aimed at parsimony and generalizability, psychologists tend to look more narrowly at data and generate their models to describe them. The problem with the later approach is that it is not clear to what environments the data generated are exportable to (hence the 1-800 phrase in the title of their paper). The implication here is that many psychology articles should have a 1-800 call center number attached to them to allow the reader to call and ask how the results would generalize to different contexts. The approach taken by behavioral economists is to look for counterexamples to rational parsimonious models and therefore use the rational models as a benchmark. Erev and Greiner think that this approach cannot be successful because, as argued by Pesendorfer, the leading behavioral models are not clear about when they apply. Hence, unless behavioral models are general and not merely constructed to account for a specific counterexample, we can easily find counterexamples to the counterexample which brings us back to the starting point.

Keith Murnighan compares the approaches of economics and social psychology to experiments and reminds the reader that while economists have the advantage of employing precise mathematical models, social psychologists have been at the experimental game for a longer time and hence have a more refined approach to experimental work. Central among this is the need to replicate research and do these replications using a variety of different tools like the laboratory, the field, simulation, and so on, since there are various audiences out there asking different types of questions. He also calls for "strong inference" which is achieved by comparing several different theories in an attempt to choose between them. Finally, Murnighan laments the fact that "we tend not to be attracted to or trained in multiple methods, much less multiple disciplines; this limits our ability to conceptualize important interdisciplinary questions and to investigate key phenomena in multiple ways."

Tom Tyler and David Amodio discuss the fact that the object of study for economists and social psychologists have historically differed. While economists have been exclusively interested in behavior and use it to infer mental states and preferences, psychologists have typically been focused directly on mental states. More recently, however, economists or at least behavioral economists have started to turn their attention to mental states. This means that they will have to make themselves comfortable with a whole host of data, such as self-reports, all geared to look inside the decision maker rather than his manifested choices. In addition, Tyler and Amodio show concern over the existence of demand effects in experiments run by economists and urge economists to consider classic experimental techniques in psychology to deal with this such as unobtrusive measurement and deception, to minimize these effects. In ending, they discuss issues of construct validity and how important this is in view of the recent popularity of neuroimaging studies.

Gary Charness agrees with Murnighan that what is needed in approaching experimental economics is a wide variety of approaches and that the best tool depends on the question being asked, but no one tool is right for all environments. In particular, he argues that the lab is better for theory testing and to identify treatment effects, whereas the field captures more realistic behavior. In either case, one needs a theory to transport findings to new populations or environments.

In Theo Offerman's response to Tyler and Amodio, he questions their suggestion that deception is a useful tool for economists in order to avoid demand effects. Basically, deception is supposed to create a decoy objective for the experiments with the idea that subjects will think that the experiment is about the decoy and offer the experimenter what he wants with respect to it, but act honestly in their behavior toward the true goal of the experiment. Offerman points out, however, that the subjects may be so focused on the decoy that they fail to pay proper attention to the true object of the experiment and thereby generate misleading data. Furthermore, Offerman doubts that many subjects are capable of figuring out the purpose of the experiment or equilibrium behavior consistent with it and, hence, are unlikely to falsely behave according to it. Finally, Offerman suggests that the way economists proceed now is probably correct: Conduct an experiment where subjects are anonymous, where there is distance between the experimenter and the subject, and where subjects are paid enough to have them focus on the trade-offs presented to them in the experiment.

Part IV: The Laboratory and the Field

In recent years, starting with the Levitt and List (2007) paper, there has been a discussion about the proper role played by field and laboratory experiments. A very naive approach would be to ask, Which is better, the lab or the field?, but as Gary Charness points out in his chapter in this volume, this is like asking, Which is better, a hammer or a screw driver? Obviously the answer depends on what it is that one wishes to accomplish by running an experiment. A more refined question needs to be asked.

In Part IV of the volume we start by reprinting the original Levitt and List (2007) paper and allow it to structure the discussion by inviting a variety of scholars to comment. At the end of Part IV, Al-Ubaydli and List reply.

The first chapter in this section is by Colin Camerer, who addresses Levitt and List criticism that lab experimental findings may not generalize to the field. The chapter has three main points. First, the special concerns over generalizability of lab results are partly due to a perception that the goal of lab experiments is to permit extrapolation to particular field settings. He calls this the "policy view." An alternative view, the "scientific view," is that empirical studies, experimental or otherwise, contribute evidence to our understanding of (a) how agents behave, (b) the role of certain incentives structures, or (c) the rules that govern interactions, in general.

Second, some results from the lab are likely to be less general than others, or not carry outside of the lab as well as others. However, this is not peculiar to the lab. Some field results are also less likely to apply to other field situations. Moreover, many of the typical features of experiments that may (or may not) limit generalizability can be varied. Finally, the chapter reviews studies that directly compare field settings to lab settings with a particular focus on List (2006).

This chapter reflects Charness's idea mentioned above that one cannot judge the usefulness of lab versus field until one specifies what the goal of the experiment is. As Camerer says, if one's goal is policy, then generalizability is crucial, because the goal is extrapolation from the experiment to the population of interest. If one's goal is scientific, however, then one views all empirical studies "as contributing evidence about the general way in which agents characteristics, incentives, rules, information, endowments and payoff structure influence economic behavior." Which is "best" therefore, depends.

Glenn W. Harrison, Morten Lau, and E. Elisabet Rutström propose a holistic approach to research where theory, lab results, common sense, field data, and econometrics are all integrated into one's research tool kit. As they put it, "Any data generated by an experiment need to be interpreted jointly with considerations from theory, common sense, complementary data, econometric methods, and expected applications." They propose a research methodology where the field and the lab are complementary and are also integrated with theory and econometrics. They illustrate their approach by considering work they have done on an artifactual field experiment in Denmark. One of the points they make is that sometimes it might make more sense to jointly estimate parameters of interest (rather than to try to develop more sophisticated elicitation methods) if, in particular, some parameters are related in important ways. Because they want to use preference estimates to evaluate policies, it is important to have subjects that are representative of the population of interest. However, they use lab experiments to determine the best procedure to perform the elicitation of the parameters of interest. Put differently, there are obviously advantages to all research methods and basically what Harrison, Lau, and Rutström are saying is why not use several when appropriate.

Using two examples, the winner's curse in common-value auctions and gift exchange in experimental labor markets, Kagel highlights the ways in which laboratory experiments and field experiments can be complementary in understanding the field setting of interest. Doing so, he highlights ways in which the interpretation of results from field experiments can be misconstrued; and by using evidence from laboratory experiments and surveys of professionals in the field, he actually paints a different interpretation of those field experiments. Specifically he highlights three points. First, learning is context specific, and this extends to professionals who use rule of thumbs that have evolved to be successful in their very specific environment, but can sometimes lead the theory to mis-predict since even though their behavior is consistent in some ways with our model's predictions, the mechanisms that drive behavior are not the ones on which the theory is based. Second, a given set of results

can often be interpreted in various ways. Third, even if the laboratory environment is far removed from the details of the field setting, if the experiments isolate important factors in that field setting, it can still provide results that are relevant for the target application.

One issue at the center of the lab/field debate is whether there is a difference between the behavior of subjects in the standard experimental subject pool, undergraduate students, and more nonstandard subjects like professionals or people who make a living engaging in a task very close to the one investigated in the lab. Guillaume Fréchette looks at this question in detail and asks whether one would reach similar conclusions about a model by running an experiment using our standard subjects or professional. If a model is rejected using undergraduate students while it finds support using professionals, it clearly undermines the external validity of the result, but it does not imply that using undergraduate students is misguided? Models are written down with no particular population in mind. They simply specify an environment and a set of incentives for agents, and those incentives and the environment are what is recreated in the experiment. A failure of the predictions of the model is a failure whether the subjects were undergraduate students or professionals. If one wants to export that result or generalize it outside the lab, to a specific group, then its generalizability may be in doubt. What Fréchette does in this chapter is to review a set of papers where both students and professionals were used as subjects and looks for whether the behavior of the students and professionals differed. The papers surveyed had to satisfy certain criteria, which were that they had to include both a sample of typical experimental subjects and a sample of subjects who are professionals at the task with which the experiment is involved. They also needed to follow standard laboratory procedures. Finally, the papers included had to be theory-based in that their aim was to test the predictions of a well-specified theory.

What Fréchette finds is that of the 13 studies that he reviews, there was very little difference between the behavior of the students and professional in comparison to the predictions of the theory, and in some of the remaining ones the students actually were closer to theory than were the professionals. This chapter, then, offers support for the proposition that when testing theory, there is little to be gained by using professions and that the behavior of subjects generalizes well.

Judd Kessler and Lise Vesterlund comment on both the Levitt–List and Camerer paper. Their comment goes to the core of what we can expect from social science and how those expectations affect the discussion between the lab and the field. The issue is basically what we expect from our theories. Do we expect them to make precise point predictions, or should we be content with theories whose comparative static predictions are borne out in the field and the lab?

Kessler and Vesterlund point out that the criticism that lab results are not externally valid is mostly about quantitative predictions, but experiments do and should focus on qualitative predictions, and nobody argues that those are not externally valid. The need to focus on directional predictions is in part a byproduct

of the need to simplify reality when constructing models (empirical or theoretical). On the other hand, they dispute the claim by Camerer that experiments should not aim for external validity. They argue that even if an experiment has no specific target application outside of the lab, if the experiment is attempting at discovering general rules of behavior, the results must be relevant outside of the lab in situations that share the same key aspects. Finally, they argue that it is not necessary for the lab to mirror the circumstances of an application of interest or to have a model indicating how to map results from the lab to an application outside of the lab for qualitative results to extrapolate.

This is an important point for social science in general. The social world is too complex to expect that our theories about it will make precise predictions. However, all policy issues are first, and we would say foremost, questions of comparative statics and directional changes; hence if our theories say something about them, they should certainly be considered useful.

David Reiley agrees with many points made by Kagel, but takes a different perspective. Because context is so important to decision making, it is necessary to study decision making by professionals in their natural environment. He sees the lack of control in field experiments as less of a problem and more of a consequence of the fact that naturally occurring economic interactions involve much more than what is assumed in economic models. Finally, because he wants to measure the importance of specific phenomena, documenting that something happens (as opposed to some theory) is insufficient. With respect to Harrison, Lau, and Rustrom, he first describes why he disagrees with their terminology defining field experiments, laboratory experiments, and so on. However, he finds the method they use for eliciting and estimating risk preferences to be an important improvement over prior methods. On the other hand, he is unsure that one can take such measurements seriously to extrapolate outside of the application in which they were recovered.

The book concludes with comments by Omar Al-Ubaydli and John List. In their chapter, Al-Ubaydli and List first discuss issues of identification, selection, and generalizability across a variety of empirical approaches. They then respond to the discussion of the Levitt and List (2007) paper presented in this section.

PART I

INTRODUCTION

CHAPTER 1

IS EXPERIMENTAL ECONOMICS LIVING UP TO ITS PROMISE?

ALVIN E. ROTH

Introduction

The question that is the title of this essay already suggests that experimental economics has at least reached a sufficient state of maturity that we can try to take stock of its progress and consider how that progress matches the anticipations we may have had for the field several decades ago, when it and we were younger. So it will help to begin by reconstructing what some of those anticipations were.

When I surveyed parts of experimental economics in Roth (1987, 1988), I hoped that experimentation would facilitate and improve three kinds of work in economics, which I called *Speaking to Theorists, Searching for Facts, and Whispering in the Ears of Princes*. By *speaking to theorists* I meant testing the empirical scope and content of theories (including especially formal theories that might depend on factors hard to observe or control outside the lab) and, in particular, testing how well and on what domains their quantitative and qualitative predictions might serve as (at least) useful approximations. By *searching for facts* I meant exploring empirical regularities that may not have been predicted by existing theories, and might even contradict them, but whose contours, once they had begun to be mapped by experiments, could form the basis for new knowledge and new theories. And by *whispering in the ears of princes* I meant formulating reliable advice, as well as communicating, justifying, and defending it.

Of course, whether experimental economics is living up to its promise could also be a question about how experimental economists are doing at developing a body of experimental methods and knowledge and creating a lively, self-sustaining, and productive community of economic research, held in high regard by the larger community. I'll address this last question first.

How Are We Doing at Building a Research Community?

There are now many experimental economists and laboratories around the world. A list compiled at the University of Montpelier locates more than 170 labs in 29 countries, including concentrations of 9 in France, 20 in Germany, 13 in Italy, 8 in Spain, 11 in the United Kingdom, and 63 in the United States.[1] There is also a professional society devoted to experimental economics, the Economic Science Association, which sponsors regular meetings, a journal, and an active internet discussion list (esa-discuss@googlegroups.com) that allows participants to quickly query the larger community with questions of all sorts, including questions about prior work and about solutions to particular problems of experimental design or analysis.

Regarding acceptance by the larger community of economists, experiments are now regularly reported in the top general-interest journals, experimenters are employed by highly ranked departments, and the Nobel memorial prize in economics has been awarded to a number of experimenters since its inception in 1968. A look at some of the Nobelists can serve as a metaphor for the progress of the field.

Maurice Allais won the prize in 1988. Although he is well known for the hypothetical choice experiment called the Allais paradox (Allais, 1953), experiments were not a persistent part of his work, nor part of the work that the Nobel committee cited him for, on general equilibrium theory. Reinhard Selten won the prize in 1994. Experiments are a large and continuous part of his work, and indeed he's one of the earliest experimenters who conducted experiments throughout his career (see, e.g., Sauermann and Selten (1959) and Selten and Chmura (2008)). But his experimental work isn't particularly related to the work he was cited for, which was the development of the theory of perfect equilibrium, a concept that in fact often performs poorly in experiments. Daniel Kahneman and Vernon Smith shared the prize in 2002. Their prize was specifically for experimental economics, and experimentation constitutes the lion's share of their work. This brings us to the 2009 prize to Elinor Ostrom. Experiments are an important part, but not the most important part, of the work she was cited for. Her experiments complement her field work (see, e.g., Ostrom (1998) and Dietz et al. (2003).[2] Much the same could be said about the 2012 prize.

There is a sense in which this history of Nobel prizes parallels how experimental economics has grown. Early experiments were done only sporadically (e.g., the Prisoner's dilemma game, which has spawned as many experiments as anything in the social sciences, was formulated for an experiment concerning Nash equilibria conducted in 1950 at the Rand Corporation and was subsequently reported by Flood (1952, 1958), who did not persist in doing experiments). Experiments became increasingly important to a relatively small group of specialists, but only more slowly achieved widespread recognition in the profession. And today experiments are flourishing in part because of how well they complement, and are complemented by, other kinds of economic research.

There are respects in which experimental economics as a community is still struggling. Chief among these is that the majority of economics departments do not have an experimental economist on their faculty, let alone a dedicated laboratory. (This may reflect low supply as much as low demand.) How this is likely to change is yet to be seen: maybe more economists will devote themselves primarily to experiments, or maybe more economists who don't primarily identify themselves as experimenters will occasionally do experiments when they need to. In the meantime, while the experimental revolution in economics is pretty well won in the journals, it still has some way to go as measured by employment in economics departments.

How Have We Done at Speaking to Theorists and Searching for Facts?

Moving back to substantive questions, how well has experimental economics begun to live up to its promise for changing the way theories are discussed, tested, and proposed? How productive has been the algorithm embodied in what we might call *the experimental cycle of creative destruction* that proceeds from Theory Testing to Exploring Unpredicted Regularities to Theory Building and back to Theory Testing?

Of course, the same "algorithm" could be said to apply to any program of theory testing, but experiments speed it up. If the test of theories waits on appropriate data to emerge through processes uncontrolled by the investigators, progress will be slow; and in economics, slow progress sometimes takes generations. But if you think that some theory or some experiment is unreliable, or is being overgeneralized or misinterpreted, you can quickly do an experiment motivated by your hypothesis, so the conversations among economists that are conducted with the help of experiments can move (relatively) fast. To my eye, this process has been quite productive, and I'll illustrate what I mean by reviewing (briefly and from a very high altitude) two experimental programs that began with theory testing, having to do with individual choice, and bargaining.

Individual Choice Behavior

I won't discuss here the long process that led to utility theory and then subjective expected utility theory becoming the canonical models of individual choice in economics, except to recall that there were a few early experiments that played a role. See Roth (1993 or 1995a) on the early history of experimental economics for accounts of (a) Bernoulli's 1738 hypothetical choice experiment on the Petersburg Paradox and (b) Thurstone's 1931 experiment on indifference curves. Even this line of experiments elicited a methodological critique of experiments in economics, by Allen Wallis and Friedman (1942).

But once utility theory had taken pride of place among economists' models of individual choice behavior, it started to be tested by experiments, including several early ones published in economics journals that foreshadowed the more sustained investigations that took place in the 1970s by psychologists as well as economists. Examples of these early studies include not only Allais' (1953) "paradoxical" demonstration of what later came to be generalized and called the common ratio effect (see, e.g., Camerer (1995)), as well as Ellsberg's (1961) demonstration of ambiguity aversion, but also demonstrations that even ordinal preferences could be intransitive (see, e.g., May (1954)).

In the 1970s, as choice experiments found yet more reproducible violations of the predictions of expected utility theory, some of the unpredicted regularities also came to be more thoroughly explored by multiple investigators, among whom Amos Tversky and Danny Kahneman were prominent. Their proposal for Prospect Theory (Kahneman and Tversky, 1979) was meant to offer a replacement for expected utility theory that captured some of these regularly observed departures from utility theory, such as the overestimation of small probabilities, as well as other behavioral regularities that were not necessarily in conflict with utility theory, such as dependence of choices on reference points, and different patterns of risk aversion and "loss aversion" with respect to gains and losses as measured against those reference points.

Together with subsequent refinements of the theory meant to allow it to be used to make predictions (Tversky and Kahneman, 1992), prospect theory came to seem to some investigators like a plausible replacement for utility theory. That is, prospect theory has a lot in common with utility (e.g., in treating individuals as having preferences), but adds parameters meant to allow it to accommodate reproducible violations of utility theory and other regularities that had been observed in experiments. [I pass silently over experiments that raised more fundamental questions about when modeling individuals as having well-defined and stable preferences is a useful approximation.]

Contemporary experiments have started focusing the sort of critical attention on prospect theory that early experiments focused on utility theory. For example, some of the regularities encoded by prospect theory turn out to be sensitive to how

choices are elicited. Harbaugh, Krause, and Vesterlund (2010) find that asking subjects to name a price they are willing to pay for a lottery induces quite different patterns of risk aversion for gambles over large and small gains and losses than does asking them to choose between a lottery and a riskless payment. In a similar way, Ert and Erev (2009) find that patterns that were interpreted as loss aversion when subjects were asked if they wished to participate in lotteries involving potential gains and losses disappear when subjects are instead asked to choose between the lottery or receiving zero for certain.

In a different kind of investigation of prospect theory, a series of papers by Ido Erev and colleagues (e.g., Barron and Erev, 2003, Erev and Barron 2005; Hertwig et al., 2004) show that how subjects react to small probabilities depends on how they learn them.[3] When subjects have lotteries described to them with numerical probabilities (as in the experiments of Kahneman and Tversky that motivated prospect theory), they tend to over-weight small probabilities. However, when subjects learn about lotteries by experiencing them multiple times (but without having the probabilities described), they tend to under-weight small probabilities.

Thus some of the behavioral regularities encoded in prospect theory may be more closely related to the way the experiments investigating utility theory were performed than was earlier appreciated.

To put his work in perspective, note that a major focus of mainstream behavioral economic research involved experiments designed to find and study counterexamples to rational decision theory, and specifically examples in which expected utility theory can be shown to make a false prediction. This led to a concentration of attention on situations in which utility theory makes a clear, falsifiable prediction; hence situations in which all outcomes and their probabilities are precisely described, so that there is no room for ambiguity about subjects' beliefs. One consequence of this is that decisions in environments in which utility theory does not make precise predictions received less attention. Environments in which participants are free to form their own beliefs fall into this category, since sometimes any decision is consistent with utility theory when beliefs cannot be observed or controlled. Decisions made in environments in which probabilities are not precisely described, but are left for subjects to learn from experience, are of this kind.

Once parts of the experimental cycle of creative destruction became less exclusively focused on utility theory, it was no longer an essential feature of good experimental design that the predictions of utility theory could be unambiguously determined for each task studied. So a much wider range of experiments opened up, some of them addressed to situations of great economic importance that had previously received less attention, like how subjects' choice behavior reflects their experience. The experiments concerning choices based on experience weren't designed to test utility theory (since they didn't pin down subjects' beliefs about the lotteries they experienced), but allowed the discovery of behavioral regularities that couldn't have been guessed from experiments that introduced the lotteries with numerical probabilities, see Erev and Roth (2014).

Even before debates about whether and how utility theory might be replaced or improved are resolved, we have gained robust new knowledge from this heritage of choice experiments. To the extent that utility theory is a useful approximation, it can only make it more useful to know that it is likely to be less accurate when choices involve small probabilities. (Whether small probabilities are likely to be over or under-weighted compared to the frequency of the events they determine may depend on how information about those probabilities is acquired. . . .) So, to the extent that we want to use utility theory as an approximation, we have become more aware that it is more of an approximation in some kinds of situations than in others.

Bargaining Behavior

A similar story of theory and experiments, followed by more theory and more experiments, could be told about many topics. In the case of bargaining theory the experimental cycle of creative destruction eventually gave rise to new theories that aim to organize data over a broader class of phenomena than merely bargaining.

In the 1970s, Nash's 1950 model was economists' canonical model of bargaining. A central assumption of Nash's model (and a number of related models, see Roth (1979)) was that the outcome of bargaining could be predicted from the information contained in the expected utility payoffs to the bargainers from each potential outcome. In particular, Nash's model predicted that the outcome of bargaining in otherwise symmetric situations would be determined by differences in the bargainers' risk aversion. It was tested by experiments that assumed that all subjects were risk neutral, and under this assumption some important aspects of its predictions could be rejected (see, e.g., Rapoport et al. (1977)). But economists were largely unpersuaded by these experiments, basically because of the feeling that a model whose predictions were based entirely on differences in risk aversion could not be adequately tested by assuming that there were no differences in risk aversion. To put it another way, the theory's predictions depend on knowing what the utility payoffs are, and without some controls, it might not be adequate to identify money payoffs with utility payoffs.

To test the theory in an environment that allowed unobserved risk aversion to be controlled, Roth and Malouf (1979) introduced the technique of binary lottery games, which has since been used fairly widely to test theories that depend on risk aversion as modeled by expected utility functions. In these experiments, payments are made in lottery tickets (i.e., in probability of winning a lottery), and the outcome of the lottery is binary; that is, the player always wins one of two prizes (sometimes one of the prizes is zero). An expected utility maximizer would be risk neutral in lottery tickets because expected utility is linear in probability; and so in an experiment that uses binary lottery payoffs, the predictions of a theory

that depends on payoffs measured in expected utility can be made unambiguous so that they can be tested.

Note that the use of binary lottery payoffs is to control for the predictions of the theory being tested, and not to control the behavior of the experimental subjects. The subjects of the experiment may not themselves be utility maximizers in the manner that a theory predicts, but binary lottery payoffs allow the experimenter to know exactly what it predicts, so that the observed behavior can be compared to the predictions. Note that there is good reason to expect that binary lottery payoffs would *not* influence subjects' behavior in the way it would if they were ideal expected utility maximizers, since the binary lottery design depends on the theory's linear treatment of probabilities, and, as noted above, subjects seem to treat (at least) small probabilities nonlinearly.

A quick digression is in order here, having to do with how I saw the promise of experimental economics, 30 years ago. We sent Roth and Malouf (1979) to the journal *Psychological Review* rather than to an economics journal because I had what turned out to be a mistaken idea of how the interaction between economics and psychology might develop. I thought that there might (continue to) be a division of labor, in which economists would theorize and psychologists would experiment, and that, in areas of potential mutual interest to economists and psychologists, what had kept economists from properly appreciating psychologists' experiments, and psychologists from properly appreciating economists' theories, might be lack of familiarity. So I thought that if we put an economic experiment in *Psych Review*, with an experimental design that would address economists' concerns about experimental tests of theories stated in expected utilities, psychologists would pick it up. In fact, psychologists weren't interested in testing the theories that attracted economists, or, when testing them, weren't interested in controlling for what economists regarded as plausible alternative hypotheses. So it turned out that if we wanted economics to have a robust experimental component, we would have to do experiments ourselves. (Although Roth and Malouf (1979) eventually became well cited by economists, I don't think it ever appealed to psychologists.)

In any event, that bargaining experiment, and some subsequent ones published in economics journals (Roth and Murnighan, 1982; Roth and Schoumaker, 1983) helped remove Nash's model from its position as economists' default model of bargaining. For a time at least, Rubinstein's (1982) alternating offer/perfect equilibrium model of bargaining over a shrinking pie became the new most popular model, with impatience (and first mover status) rather than risk aversion playing the lead role in differentiating between bargainers. And, as in the case of individual choice theory, once parts of the experimental cycle of creative destruction became less focused on bargaining theories based on expected utility theory, it was no longer an essential feature of good experimental design that the predictions of utility theory could be unambiguously determined for each task studied. Instead, tests of bargaining theory came to involve tests of perfect equilibrium, initially under the assumption that monetary payoffs were a good proxy for players' ordinal preferences.

But perfect equilibrium behavior was not observed in one-period ultimatum games (Guth et al., 1982), nor was it observed (after some preliminary suggestion otherwise by Binmore et al., 1985, 1988) in multi-period alternating offer bargaining games (Neelin et al., 1988). Ochs and Roth (1989) observed that the perfect equilibrium predictions (under the assumption that money was a good proxy for bargainers' ordinal preferences) did not do well, but also that bargainers' preferences seemed to be more complex. In particular, we noted that in our experiment and in the earlier experiments reported by others, unequal offers were often rejected and then followed by "disadvantageous counterproposals" that, if accepted, would give the bargainer a smaller monetary payoff than if he had accepted the original offer. This seemed to indicate a preference for more equal divisions, even at the cost of lower monetary payoffs.

Bolton (1991) followed with an experiment and a theory of bargaining that explicitly incorporated a preference for "fairness" in bargainers' utility functions, and this was followed by more comprehensive theories of other-regarding preferences meant to explain not only bargaining games but also market games run under comparable conditions, in which the "pecuniary" perfect equilibrium (in terms of self-regarding monetary payoffs) performed better (see Roth et al. (1991)). These new theories (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) kept the perfect equilibrium assumption, but replaced self-regarding utility functions with utility functions in which agents cared about the distribution of monetary payoffs among all players. (A different direction was taken by theories of learning proposed to explain the same kinds of experimental behavior, by replacing the perfect equilibrium assumption with a model of bounded rationality, while keeping the assumption that players were concerned only with their own payoffs; see Roth and Erev (1995).)

Each of these theories has inspired new experiments and new theories (e.g., learning theories with more or different parameters, along with other-regarding preference theories that incorporate preferences concerning intentions and expectations as well as payoffs), and these theories are in turn being tested by further experiments. So here, too, the creative cycle of experimental destruction is operating with vigor.

How Are We Doing at Whispering in the Ears of Princes?[4]

Economists give all sorts of advice, and so we whisper in the ears of many different kinds of princes. I'll focus here on market design, an area in which, recently, economists have succeeded in participating from the initial conception and design of markets all the way to their eventual adoption and implementation.

Experiments are a very natural part of market design, not least because to run an experiment, an experimenter must specify how transactions may be carried out, and so experimenters are of necessity engaged in market design in the laboratory. And experiments can serve multiple purposes in the design of markets outside of the lab. In addition to the ordinary scientific uses of experiments to test hypotheses, experiments can be used as testbeds to get a first look at market designs that may not yet exist outside of the laboratory (cf. Plott (1987)), and experiments can be used as demonstrations and proofs of concept.

My sense is that, in the first attempts to employ experiments in practical market design, experimental studies by themselves were expected to bear almost all of the weight of the argument for a particular design. More recently, experiments have served a more modest but effective role, as complements to other kinds of work, in bringing market designs from conception to implementation.

An early example of very creative experimental work with market design as its goal took place in the 1970s and 1980s, when the topic of allocating slots at the nation's busiest airports was raised. Experiments by Grether et al. (1979) and by Rassenti et al. (1982) helped direct attention to important design issues about the auction of takeoff and landing slots that remain relevant today. But they remain relevant today partly because these early efforts were unsuccessful at persuading policymakers to adopt auctions, and the political and other problems involved in doing so have yet to be solved.

Another effort to have experimental results translated directly into design decisions came in the early 1990s, when Congress passed legislation requiring the Federal Communications Commission (FCC) to design and run an auction for radio spectrum licenses. The FCC called for extensive public comment and discussion, and many economists were hired by telecommunications firms and the FCC itself to participate in the process. This eventually became one of the success stories for the involvement of economists from the initial design to implementation of a market. Plott (1997) gives an account of some of the ways that experiments and experimenters played a role in that process; see also Ledyard et al. (1997). But many of the most influential economists in the process were not experimenters, but rather auction theorists such as Paul Milgrom and Robert Wilson.

Vernon Smith (2008), in a chapter on the FCC auctions, attributes what he feels is a lack of success by experimenters at influencing policy to mistaken positions taken by policymakers and these other economists due to "entrenched resistance" (p. 131), "casual empiricism" (p. 139), "mistakes" (p. 139), "elementary errors" (p. 140), "remarkably casual empiricism" (p. 145), "early designers were all inexperienced" (p. 148), and "both users and designers have become accustomed to the fantasy that strategizing can be controlled by ever more complex rules without significantly increasing implementation costs for everyone"(p. 148).

I think that Smith may underestimate the influence that experiments had in helping to shape some of the discussions about the design of the FCC auctions,

but he is certainly correct that none of the particular design proposals advanced by experimenters were adopted.[5]

But the design of spectrum auctions is an ongoing process (although design changes now come slowly), and experiments continue to play a role in the discussion in the scientific literature, see, for example, Kagel et al. (2010) or Brunner et al. (2010) for contemporary discussions of combinatorial auctions as compared to the simultaneous ascending auctions that have become the standard design. Kagel et al. point in particular to how the development of appropriate theory helps in the design of an experiment investigating a domain like combinatorial auctions, in which the space of potential combinations and valuations created by even the simplest experimental environment is much bigger than can be meaningfully explored without some guidance about where to look.

In general, experiments have begun to play a more modest but more effective role in helping market designs by economists become implemented in functioning markets. I've worked on the design of labor markets for doctors, school choice systems, and kidney exchange, and I'll concentrate here on the design of medical labor markets, since experiments have so far been best integrated in that work (see Roth, 2002, 2008b). In particular, experiments have played roles in diagnosing and understanding market failures and successes, in exploring new market designs, and in communicating results to policy makers.

I'll briefly give examples from the design process for two medical labor markets. The first is the redesign of the labor clearinghouse through which American doctors get their first jobs, the National Resident Matching Program (see Roth and Peranson (1999)), and the second involves the reorganization of a labor market for older physicians seeking gastroenterology fellowships, the entry level positions in that subspecialty (see Niederle and Roth (2010)).

Designing Labor Markets for Doctors

New Medical Graduates

By the time I was asked in 1995 to direct the redesign of the big American clearinghouse that places most doctors in their first jobs, the National Resident Matching Program had been in operation for almost half a century, and I had studied it, as well as related clearinghouses around the world, both empirically and theoretically. The body of theory that seemed most relevant to the redesign of the NRMP was the theory of *stable matchings* (summarized at the time in Roth and Sotomayor, 1990), since Roth (1984) had shown that the early success of the NRMP in the 1950s arose when it adopted a clearinghouse that produced matchings that were stable in the

Table 1.1. Gastroenterology Markets

Market	Stable	Still in Use (Halted Unraveling)
NRMP	yes	yes (new design in 1998)
Edinburgh (1969)	yes	yes
Cardiff	yes	yes
Birmingham	no	no
Edinburgh (1967)	no	no
Newcastle	no	no
Sheffield	no	no
Cambridge	no	yes
London hospital	no	yes
Medical specialties	yes	yes (∼30 markets, 1 failure)
Canadian lawyers	yes	yes (Alberta), no (British Columbia, Ontario)
Dental residencies	yes	yes (5), no (2)
Osteopaths ($<$ 1994)	no	no
Osteopaths ($\geq$ 1994)	yes	yes
Pharmacists	yes	yes
Reform rabbis[a]	yes	yes
Clinical psychology[b]	yes	yes
Lab experiments	yes	yes
Lab experiments	no	no

[a]First used in 1997–1998.
[b]First used in 1999.

sense of Gale and Shapley (1962). Subsequent studies suggested that the stability of the outcomes played an important role in the success of other labor market clearinghouses (see, e.g., Roth (1990, 1991, 2008a)). Except for the last two lines of Table 1.1, which concern the experiment I'll come to in a moment, the table reports some of the relevant field observations. For each of the clearinghouses listed, the first column of the table reports whether it produced a stable outcome, and the second column reports whether the clearinghouse succeeded and is still in use.

From the empirical observations, stability looks like an important feature of a centralized labor market clearinghouse. Because the clearinghouses involved are computerized, their rules are defined with unusual precision, which makes questions about stability much easier to answer than in decentralized markets. Nevertheless, the empirical evidence is far from completely clear, not least because there are other differences between these markets than how their clearinghouses are organized. For example, there are differences between Edinburgh, in Scotland, and Newcastle, in England, other than whether their medical graduates were matched using a stable matching mechanism.

There are even more differences between the markets faced by medical graduates looking for jobs in Britain's National Health Service and those faced by new American doctors seeking employment in the decentralized U.S. market. The differences between those markets were very clear to American medical administrators, who therefore had reason to question whether the evidence from the British markets was highly relevant for the redesign of the American clearinghouse. And the question of whether a successful clearinghouse had to produce stable matchings had important policy implications, concerning, for example, whether the shortage of young doctors at rural hospitals could be addressed by the redesign of the clearinghouse (Roth (1986) showed that underfilled hospitals would be matched to the same set of new doctors at every stable matching).

There was thus a need for experiments to help investigate if the difference between matching mechanisms could account for the differential success of clearinghouses that had been observed to fail or to succeed in the field. That is, an experiment would allow these different mechanisms to be examined without the confounding effect of differences between different regions of the British National Health Service, for example.

Kagel and Roth (2000) reported an experiment that compared the stable algorithm used in Edinburgh and Cardiff with the unstable "priority" algorithm used in Newcastle and in slightly different versions in Birmingham and Sheffield. The point of the experiment was not, of course, to reproduce the field environments, but rather to create a simpler, more controlled environment in which the clearinghouse algorithm could be changed without changing anything else.

The experiment examined laboratory markets consisting of six firms and six workers (half "high productivity," half "low productivity"). Subjects received about \$15 if they matched to a high productivity partner, and around \$5 if they matched to a low productivity partner, and there were three periods in which matches could be made: –2, –1, 0, with the final payoff being the value of the match minus \$2 if made in period –2, or minus \$1 if made in period –1. That is, there was a cost for matching early, before period 0.

However, the experimental markets initially offered only a decentralized match technology: Firms could make one offer in any period if they were not already matched. Workers could accept at most one offer. This decentralized matching technology suffers from congestion: Firms would like to make more offers than they are able to at period 0, and a firm that waited until period 0 to make an offer would run a risk that its offer would be refused by a worker who had received a preferable offer, and it would be unmatched. So firms learned from experience that they had to make offers early, even though this was costly. (In this simple experiment, the costs of going early were simply the fines imposed by the experimenters.)

After experiencing 10 markets using this decentralized technology, a centralized matching technology was introduced for period 0 (periods –2 and –1 were organized as before). Participants who were still unmatched at period 0 would submit rank order preference lists to a centralized matching algorithm.

The experimental variable was that the matching algorithm would be either (a) the unstable priority algorithm used in Newcastle or (b) the stable matching algorithm used in Edinburgh.

The experimental results reproduce what we see in the field: The stable matching mechanism reverses unraveling, whereas the unstable one does not. In addition, the experiment allows us to observe more than the data from the field. We can see not only who matches to whom, but also the pattern of offers and acceptances and rejections, which turns out to be quite revealing. In particular, the introduction of the stable matching mechanism, which reversed the unraveling, did so not by making firms unwilling to make early offers, but by making it safe for workers to decline them. This experimental observation was confirmed in subsequent field and experimental studies of the market for lawyers (see Avery et al. (2001, 2007) and Haruvy et al. (2006)) and played a role in the subsequent design of the gastroenterology labor market described below.

Note how the laboratory experiments fit in Table 1.1's list of observations, and also note how they complement the variety of matching mechanisms observed in the field. The lab observations are by far the smallest but most controlled of the markets on the list (which otherwise range over two orders of magnitude in size, from the large American market for new doctors, which fills more than 20,000 positions a year, to the smallest British markets and American fellowship markets, some of which fill fewer than 100 positions a year). The laboratory markets also offer the smallest incentives, far smaller than the career-shaping effects of a first job.

So, by themselves, the laboratory experiments would likely not be seen as providing strong evidence that the large American medical clearinghouse needed to produce stable matchings. But, by themselves, the field observations left open the possibility that the success and failure of the various clearinghouses is unaffected by the stability of the matching mechanism and that the apparent connection is only coincidental. The field observations also leave open the possibility that the experience of the British markets in this regard depends in some way on the complex ways in which British medical employment differs from that in the United States.

Taken together, the field evidence plus the laboratory evidence give a much clearer picture. In the laboratory experiments, the success of the stable mechanism and the failure of the unstable mechanism can be unequivocally attributed to the difference between the two mechanisms, since, in the lab, the markets are controlled so that this is the only difference between them. The laboratory outcomes thus add weight to the hypothesis that this difference is what caused the same outcomes in the field, in Edinburgh and Newcastle, even though there are other differences between those two cities. And seeing this effect in the simple laboratory environment shows that the choice of algorithm has an effect that is not simply a function of some of the complexities of the British medical market. Together with the large body of theoretical knowledge about stable mechanisms, the laboratory experiment and field observations thus provided quite helpful guidance

about how the redesign of the clearinghouse should proceed, and they supported the hypothesis that stability is an important ingredient of a successful labor market clearinghouse of this kind. The current NRMP clearinghouse employs the stable Roth and Peranson (1999) algorithm.

So the experiments fit very naturally on the list of markets studied in Table 1.1. They are the smallest but clearest, and they illuminate and are illuminated by the similar results observed in the larger, naturally occurring markets on that list.[6]

Gastroenterology Fellows

In a similar way, helping gastroenterologists redesign the labor market for new gastroenterology fellows in 2006 required a mix of field and experimental studies.

A gastroenterology fellowship is the entry-level job for the internal medicine subspecialty of gastroenterology, and doctors can take this position after they have become board-certified internists by completing a three-year residency in internal medicine. So, when the gastroenterology labor market started to unravel in the 1980s, gastroenterologists were already familiar with labor market clearinghouses, since they had all participated in the resident match, the NRMP. A fellowship match program was set up in 1986, but in 1996 it suddenly began to fail, and soon completely collapsed, with fellowship programs once again hiring fellows outside of the match.

There was considerable disagreement about the cause of this failure, and a combination of field studies and an experiment helped clarify this (see Niederle and Roth (2003, 2004) and McKinney et al. (2005)). The field evidence consisted of one set of observations of a complex historical event leading to the failure of the clearinghouse, which was consistent with many hypotheses. These could be investigated in laboratory attempts to make a clearinghouse fail under similar circumstances. To make a long story short, part of what happened in 1996 is that there was an announced and widely anticipated reduction in the number of fellowship positions (together with an increase from two to three years needed to become a board-certified gastroenterologist). This reduction in the number of positions was accompanied by an even larger and unexpected reduction in the number of doctors applying for those positions. As it happened, despite the reduction in the number of positions, 1996 turned out to be the first year in which the number of positions exceeded the number of applicants. It now appears that the collapse of the clearinghouse began when fellowship programs (alarmed by the smaller than expected number of applicants they received) made early offers to applicants, who accepted them without waiting for a match.

Of course, there are other ways the historical story could be parsed. But McKinney et al. (2005) found in the laboratory that anticipated shifts in supply and demand, visible to both sides of the market, did not cause declines in match participation anywhere near the magnitude caused by unanticipated shocks, particularly

when these are more visible to one side of the market than to the other.[7] In particular, we looked at shifts in demand that were visible either to both firms and workers or only to firms (as when an unexpected change in demand is visible to firms who receive few applications, but not to workers). Demand reductions of both kinds caused firms to try to make more early hires, but when workers knew that they were on the short side of the market they were more likely to decline such offers than when they were unaware of the shift in demand. In the lab it was clearly the combination of (a) firms making early offers outside of the match and (b) workers not feeling safe to reject them and wait for the match that caused the market to unravel. The experimental results also clearly suggested that, after such a shock, it would be possible to reestablish a functioning match.

This experiment, like that of Kagel and Roth (2000), also suggested that when there was not much participation in the match, there would be pressure for the market to unravel, with participants making offers earlier and earlier. But this is an observation that is clearly built into the experimental design, almost as an assumption, since in the experiment, early offers were one of very few strategic options available. So the experiment by itself didn't provide much evidence that unraveling was going on in the gastroenterology market. Establishing this depended on field data, both from employer surveys and analysis of employment data, which showed that, 10 years after the collapse of the match, the market continued to unravel, with employers making exploding offers earlier each year than the previous year, not all at the same time, and months ahead of the former match date (Niederle et al., 2006). This also had the consequence of causing a formerly national market to have contracted into much more local, regional markets.

Taken together, the field and experimental evidence made what proved to be a convincing case that the absence of a match was harmful to the market and that the collapse following the events of 1996 had been due to a particular set of shocks that did not preclude the successful operation of a clearinghouse once more.

But a problem remained before a clearinghouse could be restarted. The employers were accustomed to making early exploding offers; and program directors who wished to participate in the match worried that if their competitors made early offers, then applicants would lose confidence that the match would work and consequently would accept those early offers, because that had been the practice in the decentralized market. That is, in the first year of a match, applicants might not yet feel that it is safe to reject an early offer to wait for the match. Program directors who worried about their competitors might thus be more inclined to make early, pre-match offers themselves.

There are decentralized markets that have avoided the problem of early exploding offers, in ways that seemed to suggest policies that might be adopted by the professional gastroenterology organizations. One example is the market for Ph.D. students, in which a policy of the Council of Graduate Schools (adopted by the large majority of universities) states that offers of admission and financial

support to graduate students should remain open until April 15. Specifically, the policy states in part:

> Students are under no obligation to respond to offers of financial support prior to April 15; earlier deadlines for acceptance of such offers violate the intent of this Resolution. In those instances in which a student accepts an offer before April 15, and subsequently desires to withdraw that acceptance, the student may submit in writing a resignation of the appointment at any time through April 15.

This of course makes early exploding offers much less profitable. A program that might be inclined to insist on an against-the-rules early response is discouraged from doing so in two ways. First, the chance of actually enrolling a student who is pressured in this way is diminished, because the student is not prevented from later receiving and accepting a more preferred offer. Second, a program that has pressured a student to accept an early offer cannot offer that position to another student until after the early acceptance has been declined, at which point most of the students in the market may have made binding agreements. In the market for new Ph.D. students, this policy has helped to make early exploding offers a non-issue.

But gastroenterologists were quick to point out that there are many differences between gastroenterology fellowships for board-certified internists and graduate admissions for aspiring PhDs. Perhaps the effectiveness of the CGS policy depended in some subtle way on the many and complex differences between these two markets. So experiments still had another role to play before a marketplace could be built that would reverse the previous decade of unraveling. And here the role of experiments was (once again) to help bridge the gap, in the laboratory, between two rather different markets, namely, the gastroenterology market and the market for admissions of Ph.D. students to graduate programs. (Recall our earlier discussion of the differences between British and American markets for new doctors.)

Niederle and Roth (2009) bridged this gap by studying in a simple laboratory environment the effect of the CGS policy of empowering students to accept offers made before a certain time and then change their minds if they received offers they preferred.[8] In the lab, early inefficient matches that were common when subjects could not change their minds about early offers and acceptances essentially disappeared when the policy allowing changes of mind was in place. And the rise in efficiency came about not because early offers were made, accepted, and then subsequently rejected, but rather because this possibility discouraged early offers from being made. The fact that this could be observed in the transparently simple laboratory environment showed that the policy did not depend for its effectiveness on some subtle feature of the complex Ph.D. admissions process.

The four gastroenterology organizations adopted the policy, as proposed in Niederle et al. (2006), and the gastroenterology match for 2007 fellows was held June 21, 2006. It succeeded in attracting 121 of the 154 eligible fellowship programs (79%). 98% of the positions offered in the match were filled through the match.

Niederle et al. (2008) show that in the second year of the new centralized match the interview dates were successfully pushed back and are now comparable to those of other internal medicine specialties that have used a centralized match for many years.

To summarize the role of experiments in practical market design, simple experiments can help us understand complex markets. They fill a gap left even by field studies of similar markets, since comparison between one complicated market and another is often quite properly viewed with suspicion; for example, the market for new American doctors is indeed very different from the comparable markets for British doctors, and the market for gastroenterology fellows is very different from the market for Ph.D. students.

Of course, experiments are also very different from complex markets like those for doctors, but they are different in simple, transparent ways. As such, they can sometimes have more *ecological validity* than observations from natural markets that are equally as complex as, but different from, a particular market. By the same token, evidence about complex markets drawn entirely from simple experiments is also likely to be less convincing than a wide array of observations from markets of different transparency and complexity.

Experimental economics has thus become more effective and important for practical market design as it has become less "heroic" and stand-alone. Experiments (in the lab or in the field) seem to have the greatest effect when they are used together with other kinds of investigations, both observational and theoretical.

How Are We Doing at Generating Productive New Areas of Research?

One question that seems natural for diagnosing the current state of experimental economics is whether it is continuing to generate new areas of investigation. Certainly the field has been fertile in the last few decades; for example, experiments are at the heart of the great growth in study of what is often called behavioral economics, but might better be described as Economics and Psychology (E&P). This line of work includes not only the venerable study of failures of utility theory and of heuristics and biases, in the style of Kahneman and Tversky, but also some very new components, such as the emerging "neuroeconomics" conversation between experimental economists and neuroscientists.[9]

Loosely speaking, the main project in E&P has been to better understand individuals' thought processes as they make choices. The attraction of neuroeconomics, with tools like fMRI scanners, is to find real-time correlates of these choice processes in the brain. Time will tell how productive this kind of research will be for economists' agenda (a subject on which there has been much discussion, to which I will return briefly in the next section).

But I take it as a sign of the healthy state of experimental economics that there are other, related but different research programs underway that seem to be growing quickly. One of them is the emerging study of what might be called Economics and Biology (E&B). It, too, has historical antecedents,[10] and, like E&P (which purists might want to argue is a subset of E&B), it has many parts, including concerns with nutrition and disease. But for comparison purposes, it may be helpful to focus on a topic shared by E&P and E&B, namely, individual choices and preferences. One difference between the two approaches is that E&P has a strong emphasis on thought processes *while* making a choice, whereas E&B directs our attention to some of the *longer-term determinants* of individuals' choices, actions, and preferences.

At the level of the whole human organism are studies of gender and how men and women may behave differently in economic environments. For example, the experimental studies of Gneezy et al. (2003) and Niederle and Vesterlund (2007) address the question of whether men and women behave differently in competitive environments, and whether they make different choices about whether to engage in competition.

Closer to the level of brain chemistry, a small literature is growing up around "endocrinological economics," having to do with the effect of hormones on behavior, including, for example, studies of risk-taking behavior and the menstrual cycle, but also studies of how exposure to testosterone in the womb (e.g., as measured by the ratio of the length of the second and fourth fingers) is correlated with risk-taking behavior of adult men and women.[11] A related line of work measures the *heritability* of preferences such as risk-taking propensities—for example, through studies of twins (see, e.g., Cesarini et al. (2009)).

It will likely be quite some time before the impact of this developing literature on economics, or on psychology and biology, can be assessed. But the growing collaboration of economists with other kinds of scientists (such as neuroscientists and biologists generally) interested in using economic experiments to gain insight into aspects of human biology is a sign of how useful and flexible the experimental economics laboratory is. And the fact that experimental economics is full of new research directions, of which this is just one example, is an indication of the thriving life of experimentation within economics.

Critiques and Criticisms of Experiments (from Within and Without)

Perhaps one of the clearest signals that experimental economics is coming of age is that we have recently seen (a) a wide variety of commentaries, criticisms, and critiques, both from those who do experiments and recommend them and from those who don't and don't; (b) some crossovers; and (c) comments from those who favor

one style of experimental research over another.[12] Only recently has experimental economics become so widely perceived as successful that it can be in a position to experience any sort of *backlash*.[13] Experimental economics is even approaching the decadent phase in which it becomes the subject of philosophy of science (see, e.g., Guala (2005) and Bardsley et al. (2010), not to mention conferences and volumes on its methods and vital signs).[14]

The tone of the criticisms and critiques ranges from temperate to hot, from thoughtful to polemic, sometimes in the same article. Like articles reporting experiments, articles criticizing or praising them can suffer if the conclusions are too broad, even if they also contain valid points. Nevertheless, there is always the potential opportunity to learn something about our craft, and its place in economics, by paying some attention to the criticisms of experiments and to the arguments made in their defense.

Most of the critiques touch on both a methodological theme and a substantive one. See, for example, the discussion of neuroeconomics by Bernheim (2009), Gul and Presendorfer (2008), Rustichini (2009), and Sobel (2009), or the discussions of inequity aversion and experimental methods and reporting by Binmore and Shaked (2010a, b), Fehr and Schmidt (2010), and Eckel and Gintis (2010).

One of the most broadly controversial lines of criticism among experimenters has to do with the comparative advantages of experiments conducted in the laboratory and in the field. The papers by Levitt and List (2007a,b) are widely read as proposing that laboratory experiments are in some important senses simply inferior to field experiments, particularly regarding their *ecological validity*—that is, the ability of their conclusions to generalize to naturally occurring target markets. In my remarks above on market design I've already made clear that I don't agree, although I certainly think that studies in the lab and in the field (whether experimental or not) can complement each other.[15]

It's worth noting in this regard that field experiments have come in for some criticism of the same sort from applied econometricians; see, for example, the recent papers by Deaton (2009) and by Heckman and Urzua (2010), as well as the reply by Imbens (2010), all of which discuss the extent to which the local average treatment effects that are revealed by randomized field experiments can be appropriately generalized from one domain to another. In this connection, see also Falk and Heckman (2009), whose paper's title conveys its central message "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." Their concluding paragraph is one with which I agree:

> "Causal knowledge requires controlled variation. In recent years, social scientists have hotly debated which form of controlled variation is most informative. This discussion is fruitful and will continue. In this context it is important to acknowledge that empirical methods and data sources are complements, not substitutes. Field data, survey data, and experiments, both lab and field, as well as standard econometric methods, can all improve the state of knowledge in the social sciences. There is no hierarchy among these methods and the issue of generalizability of results is universal to all of them."

I suspect that the reason they found it necessary to argue in favor of what should be such an uncontroversial conclusion is that these methodological debates have been conducted with much more heat than are the usual scientific disagreements. They sometimes have the flavor of accusation, as if those who disagree must be bad scientists who ignore critical evidence. I'm puzzled by this, but I think it may have something to do with the general problem of drawing conclusions from evidence, in a way that is made particularly stark by experimental evidence.

I've elaborated on this for many years in my experimental economics classes with the following dramatization designed to show how different people can view the same evidence differently.[16]

> Suppose I show you what appears to be a deck of playing cards, all face down, and propose that we investigate the hypothesis that none of the cards is blue, by turning them face up one at a time. The first 20 cards are all red and black, Hearts and Clubs and Spades and Diamonds. After each one is turned up, we both agree that our confidence that none of the cards is blue has increased. The 21st card is the four of Forest, which is, of course, green. I argue that since yet another card that isn't blue has been observed, my confidence that none of them are blue has increased. But you argue that if the deck contains the four of Forest, you suddenly realize it might contain the six of Sky, which is, of course, blue. So, on the basis of the same evidence, and even though none of the cards turned over so far is blue, your confidence in the no-blue hypothesis is diminished (because the hypothesis you really believed going in was the unstated one that we were dealing with a standard deck of playing cards, i.e. that there were only red and black cards).

The fact that we disagree based on the same evidence might be enough to disrupt our future card games, especially if we draw our conclusions with unjustified confidence or generality. But, fortunately, there is a way forward toward resolving such disagreements, which is to gather more evidence, including designing and conducting more experiments. So the recent methodological disputes will probably be good for business in the longer term.

In Conclusion

Experimental economics is thriving. It is living up to its promise, although we are also learning, by doing experiments, more about what experiments promise for economics.

Experiments are becoming better integrated with other kinds of economic research, including a vigorous conversation with theorists that motivates new theories and that also tests existing ones. We are learning more about how experiments are complements to other kinds of empirical investigation. The lab has unique advantages, as do field data and theoretical models. (Just as modern warfare cannot be won by air power alone, but no commander would voluntarily give up his air force,

experiments are often not enough on their own to carry the day, but no science should try to do without them.) Experiments are essential when control is needed.

Let's celebrate our successes and learn from our experience, but not spend too much time looking backward and inward. There's good work to be done.

NOTES

1. http://leem.lameta.univ-montp1.fr/index.php?page=liste_labos&lang=eng, accessed September 2014.

2. John Nash, who shared the 1994 prize with Selten and Harsanyi, and Thomas Schelling, who shared the 2005 prize with Aumann, also have been involved in some significant experiments, although experiments didn't play a large role in their careers. But Schelling (1957, 1958, 1960) reported important early experiments.

3. The 2003 and 2004 papers address this issue directly; the 2005 paper looks at models of learning that might help account for it.

4. This section borrows from my forthcoming chapter on Market Design in volume 2 of the *Handbook of Experimental Economics*, in which more detail will be found.

5. Milgrom (2007) adds an interesting dimension to the discussion of how experiments were used in the policy discussion, writing of one of the consulting reports (Cybernomics, 2000) that presented a particular proposal based on an experiment (p. 953): "Cybernomics presented its results to the FCC in a report and at a conference, where they were represented by two highly regarded academic experimenters: Vernon Smith and David Porter. . . . The Cybernomics report is not detailed enough to enable a fully satisfactory assessment of its results. The FCC contract did not require that detailed experimental data be turned over to the sponsors. When the FCC and I later asked for the data, we were told that they had been lost and cannot be recovered."

6. Other experiments illuminated some of the outlier results—for example, regarding the single-medical-school markets at the London Hospital and Cambridge. See Ünver (2001, 2005).

7. Subjects in the roles of workers and firms first participated in 15 three-period decentralized markets with a congested (one offer per period) match technology and a cost for matching early, then in 15 markets with the same number of firms and workers in which a centralized clearinghouse was available to those who remained unmatched until the last period, and then in 15 further markets in which a change was made in the number of either firms or workers, a change that was observable to firms but only observable to workers in some treatments.

8. Each market involved five firms and six applicants and consisted of nine periods in which firms could make offers. (So this decentralized market was not congested; that is, there was enough time for all offers to be made.) Firms and applicants had qualities, and the payoff to a matched firm and applicant was the product of their qualities. Firms' qualities (1, 2, 3, 4, and 5) were common knowledge, but applicants' qualities were stochastically determined over time: In periods 1, 4, and 7 each applicant received an integer signal from 1 to 10 (uniform iid). The quality of each applicant was determined in period 7 through the relative ranking of the sum of their three signals: The applicant with the highest sum had a quality of 6, the second highest had a quality of 5, and the lowest had a quality of 1 (ties were broken randomly). So efficient matches (which assortatively match

the applicants and firms in quality order) can only be made if matching is delayed until applicants' qualities have been determined by the final signal in period 7. But lower-quality firms have an incentive to try to make matches earlier, since this gives them their only chance at matching to higher-quality workers.

9. New opportunities for experiments come about both as new subject matter is considered and as new possibilities for running experiments arise. In just this way, neuroeconomics reflects both interest in the brain and the increasing availability of tools for measuring aspects of brain activity. Other new possibilities for running experiments are likely to be increasingly exploited in the future. The ubiquity of the internet will surely lead to increasing numbers of experiments using it, not merely on websites set up by experimenters, but also on sites set up for other purposes, like sales, social networking, labor market matching, dating, advice, and so on.

10. For example, William Stanley Jevons, better remembered today for his 1876 *Money and the Mechanism of Exchange*, reported a set of three experiments about efficient weight lifting (and throwing) in an 1870 article in *Nature* (which I first learned of in Bardsley et al., 2010). At that time, of course, a lot of economic activity was muscle-powered.

11. One indicator of the vibrant nature of the experimental community is that in August 2009 Burkhard Schipper sent out a request on the ESA-discuss chat group for papers on endocrinology and economics, and later in the month he circulated the bibliography of papers he had received, consisting of about 30 economics papers written since 2002, with titles containing "testosterone," "oxytocin," "dopamine," "second-to-fourth digit ratio". . . (see Apicella et al. (2008), Baumgartner et al. (2008), Burnham (2007), Caplin and Dean (2008), Chen et al. (2009), Coates et al. (2009), and Pearson and Schipper (2009) for a sampling of recent papers whose titles are evocative of this area.)

12. Recall from Wallis and Friedman (1942) that methodological critiques of experiments in economics aren't entirely new. But when there wasn't much experimental economics, there wasn't much to analyze, criticize, or debate. In the 1990s there was some internal discussion of methods and goals. See, for example, Roth (1994) on what was then sometimes the practice of labeling as an experiment each session of a larger experimental design, with some resulting irregularities in how experiments were reported. Today the practice of reporting whole experimental designs seems to be much more widely followed. See also some discussion of the differing emphases and interpretations that arose from attempts to distinguish between "experimental" and "behavioral" economics—for example, in Binmore (1999) and Loewenstein (1999).

13. Not so long ago, there was hardly any criticism of experimental economics; instead, there were just statements to the effect that "economics is not an experimental science." (If you google that phrase, you can still find some examples, about half of them with the original meaning, and the other half by experimenters reflecting on the change in the views in the profession.)

14. Some of the recent philosophy of economics (as it is called when addressed to philosophers of science) and methodology (when addressed also to economists) has been written by distinguished experimenters; see, for example, the special issue of the *Journal of Economic Behavior and Organization*, entitled "On the Methodology of Experimental Economics" (Rosser and Eckel, 2010), beginning with a target article by Smith (2010) and followed by many replies and a final article by Croson and Gächter (2010). See also Bardsley et al. (2010) and Starmer (1999), but also philosophical investigations of the uses of theoretical models that resonate with the use of experiments as models, such as

Sugden (2009). In this connection see also Mäki (2005). (I write as a frequently disappointed consumer of philosophy of science who nevertheless returns sporadically in the hope that it will teach me how to do science better. I am often reminded of the quip attributed to the late Richard Feynman that philosophy of science is as useful to scientists as ornithology is to birds. Needless to say, this may just mean that birds are misguided if they look to ornithology for advice; ornithology is interesting in its own right, even if not so useful to practicing birds. In this respect, it is another sign of the progress of experimental economics that it now commands some attention from those who study what economists do.)

15. See also Banerjee and Duflo (2008) and Cohen and Easterly (2009).

16. I suspect that something like this may be an old philosophy of science chestnut, but correspondence with philosophers who specialize in "confirmation theory" and related matters hasn't turned up a source, although it has turned up a number of related examples illustrating the difficulty of interpreting evidence.

REFERENCES

Allais, M. 1953. Le Comportement de L'Homme Rationnel Devant le Risque: Critique des Postulats et Axiomes de L'Ecole Americane. *Econometrica* **21**:503–546.

Apicella, C. L., A. Dreber, B. Campbell, P. B. Gray, M. Hoffman, and A. C. Little. 2008. Testosterone and Financial Risk Preferences. *Evolution and Human Behavior* **29**:384–390.

Avery, C., C. Jolls, R. A. Posner, and A. E. Roth. 2001. The Market for Federal Judicial Law Clerks. *University of Chicago Law Review* **68**:793–902.

Avery, C., C. Jolls, R. A. Posner, and A. E. Roth. 2007. The New Market for Federal Judicial Law Clerks. *University of Chicago Law Review* **74**:447–486.

Banerjee, A. V. and E. Duflo. 2008. The Experimental Approach to Development Economics. NBER Working Paper No. 14467.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden. 2010. *Experimental Economics: Rethinking the Rules*. Princeton, NJ: Princeton University Press.

Barron, G. and I. Erev. 2003. Small Feedback-Based Decisions and Their Limited Correspondence to Description Based Decisions. *Journal of Behavioral Decision Making* **16**:215–233.

Baumgartner, T., M. Heinrichs, A. Vonlanthen, U. Fischbacher, and E. Fehr. 2008. Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans. *Neuron* **58**:639–650.

Bernheim, B. D. 2009. "On the Potential of Neuroeconomics: A Critical (but Hopeful) Appraisal." *American Economic Journal: Microeconomics,* **1**(2): 1–41.

Bernoulli, D. 1738. Specimen Theoriae Novae de Mensura Sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* **5**:175–192.

Binmore, K. 1999. Why Experiment in Economics? *Economic Journal* **109**, F16–F24.

Binmore, K. and A. Shaked. 2010a. Experimental Economics: Where Next? *Journal of Economic Behavior and Organization* **73**:87–100.

Binmore, K. and A. Shaked. 2010b. Experimental Economics: Where Next? Rejoinder. *Journal of Economic Behavior and Organization* **73**:120–121.

Binmore, K., A. Shaked, and J. Sutton. 1985. Testing Noncooperative Bargaining Theory: A Preliminary Study. *American Economic Review* **75**:1178–1180.

Binmore, K., A. Shaked, and J. Sutton. 1988. A Further Test of Noncooperative Bargaining Theory: Reply. *American Economic Review* **78**:837–839.

Bolton, G. 1991. A Comparative Model of Bargaining: Theory and Evidence. *American Economic Review* **81**:1096–1136.

Bolton, G. and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity and Competition. *American Economic Review* **90**:166–193.

Brunner, C., J. K. Goeree, C. A. Holt, and J. O. Ledyard. 2010. An Experimental Test of Flexible Combinatorial Spectrum Auction Formats. *American Economic Journal: Microeconomics* **2**(1):39–57.

Burnham, T. C. 2007. High-Testosterone Men Reject Low Ultimatum Game Offers. *Proceedings of the Royal Society* **274**:2327–2330.

Camerer, C. 1995. Individual Decision Making. In *Handbook of Experimental Economics*, eds. J. H. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press, pp. 587–703.

Caplin, A. and M. Dean. 2008. Dopamine, Reward Prediction Error, and Economics. *Quarterly Journal of Economics* **123**:663–701.

Cesarini, D., C. T. Dawes, M. Johannesson, P. Lichtenstein, and B. Wallace. 2009. Genetic Variation in Preferences for Giving and Risk-Taking. *Quarterly Journal of Economics* **124**:809–842.

Chen, Y., P. Katuscak, and E. Ozdenoren. 2009. Why Can't a Woman Bid More Like a Man? Unpublished.

Coates, J. M., M. Gurnell, and A. Rustichini. 2009. Second-to-Fourth Digit Ratio Predict Success Among High-Frequency Financial Traders. *Proceedings of the National Academy of Sciences* **106**:623–628.

Cohen, J. and W. Easterly. 2009. *What Works in Development?: Thinking Big and Thinking Small*. Washington, DC: Brookings Institution Press.

Croson, R. and S. Gächter. 2010. The Science of Experimental Economics. *Journal of Economic Behavior and Organization* **73**:122–131.

Cybernomics, Inc. 2000. An Experimental Comparison of the Simultaneous Multi-Round Auction and the CRA Combinatorial Auction. Submitted to the Federal Communications Commission, http://wireless.fcc.gov/auctions/conferences/combin2000/releases/98540191.pdf

Deaton, A. 2009. Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. NBER Working Paper No. 14690.

Dietz, T, E. Ostrom, and P. C. Stern. 2003. The Struggle to Govern the Commons. *Science* **302**:1907–1912.

Eckel, C. and H. Gintis. 2010. Blaming the Messenger: Notes on the Current State of Experimental Economics. *Journal of Economic Behavior and Organization* **73**:109–119.

Ellsberg, D. 1961. Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics* **75**:643–669.

Erev, I. and G. Barron. 2005. On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychological Review* **112**(4):912–931.

Erev, I. and A. E. Roth. 2014. Maximization, Learning and Economic Behavior. *Proceedings of the National Academy of Science* **111**(3):10818–10825.

Ert, E. and I. Erev. 2009. On the Descriptive Value of Loss Aversion in Decisions Under Risk. Unpublished.

Falk, A. and J. J. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* **326**:535–538.

Fehr, E. and K. M. Schmidt. 1999. A Theory Of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* **114**:817–868

Fehr, E. and K. M. Schmidt. 2010. On Inequity Aversion: A Reply to Binmore and Shaked. *Journal of Economic Behavior and Organization* **73**:101–108.

Flood, M. M. 1952. Some Experimental Games. Research Memorandum RM-789, RAND Corporation.

Flood, M. M. 1958. Some Experimental Games. *Management Science* **5**:5–26.

Gale, D. and L. Shapley. 1962. College Admissions and the Stability of Marriage. *American Mathematical Monthly* **69**:9–15.

Gneezy, U., M. Niederle, and A. Rustichini. 2003. Performance in Competitive Environments: Gender Differences. *Quarterly Journal of Economics* **118**:1049–1074.

Grether, D. M., R. M. Isaac, and C. R. Plott. 1979. Alternative Methods of Allocating Airport Slots: Performance and Evaluation, Prepared for Civil Aeronautics Board Contract Number 79-C-73, Polinomics Research Laboraories, Inc., Pasadena, CA.

Guala, F. 2005. *The Methodology of Experimental Economics*. New York: Cambridge University Press.

Gul, F. and W. Pesendorfer. 2008. The Case for Mindless Economics [paper]. *The Foundations of Positive and Normative Economics,* eds. Andrew Caplin and Andrew Shotter. Oxford University Press.

Guth, W., R. Schmittberger, and B. Schwarz. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* **3**:367–388.

Harbaugh, W. T., K. Krause, and L. Vesterlund. 2010. The Fourfold Pattern of Risk Attitudes in Choice and Pricing Tasks. *Economic Journal*. **120**(545):569–611.

Haruvy, E., A. E. Roth, and M. U. Ünver. 2006. The Dynamics of Law Clerk Matching: An Experimental and Computational Investigation of Proposals for Reform of the Market. *Journal of Economic Dynamics and Control* **30**:457–486.

Heckman, J. and S. Urzua. 2010. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. *Journal of Econometrics* **156**(1):27–37.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. 2004. Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science* **15**:534–539.

Imbens, G. W. 2010. Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*. **48**(2):399–423.

Jevons, W. S. 1870. On the Natural Laws of Muscular Exertion. *Nature* **2**:158–160 (30 June 1870).

Kahneman, D. and A. Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* **47**:263–291.

Kagel, J. H., Y. Lien, and P. Milgrom. 2010. Ascending Prices and Package Bidding: A Theoretical and Experimental Analysis. *American Economic Journal: Microeconomics* **2**, August, 160–185.

Kagel, J. H. and A. E. Roth. 2000. The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment. *Quarterly Journal of Economics* **115**:201–235.

Ledyard, J. O., D. Porter, and A. Rangel. 1997. Experiments Testing Multiobject Allocation Mechanisms. *Journal of Economics and Management Strategy* **6**:639–675.

Levitt, S. D. and J. A. List. 2007a. What Do Laboratory Experiments Measuring Social Preferences Tell Us About the Real World. *Journal of Economic Perspectives* **21**(2):153–174.

Levitt, S. D. and J. A. List. 2007b. Viewpoint: On the Generalizability of Lab Behaviour to the Field. *Canadian Journal of Economics* **40**(2):347–370.
Loewenstein, G. 1999. Experimental Economics from the Vantage-Point of Behavioural Economics. *Economic Journal* **109**, F25–F34.
Mäki, U. 2005. Models Are Experiments, Experiments Are Models. *Journal of Economic Methodology* **12**:303–315.
May, K. O. 1954. Intransitivity, Utility, and the Aggregation of Preference Patterns. *Econometrica* **22**:1–13.
McKinney, C. N., M. Niederle, and A. E. Roth. 2005. The Collapse of a Medical Labor Clearinghouse (and Why Such Failures Are Rare). *American Economic Review* **95**:878–889.
Milgrom, P. 2007. Package Auctions and Exchanges. *Econometrica* **75**:935–965.
Nash, J. 1950. The Bargaining Problem. *Econometrica* **28**:155–162.
Neelin, J., H. Sonnenschein, and M. Spiegel. 1988. A Further Test of Noncooperative Bargaining Theory: Comment. *American Economic Review* **78**:824–836.
Niederle, M., D. D. Proctor, and A. E. Roth. 2006. What Will Be Needed for the New GI Fellowship Match to Succeed? *Gastroenterology* **130**:218–224.
Niederle, M., D. D. Proctor, and A. E. Roth. 2008. The Gastroenterology Fellowship Match — The First Two Years. *Gastroenterology* **135**, 2 (August):344–346.
Niederle, M. and A. E. Roth. 2003. Unraveling Reduces Mobility in a Labor Market: Gastroenterology with and without a Centralized Match. *Journal of Political Economy* **111**:1342–1352.
Niederle, M. and A. E. Roth. 2004. The Gastroenterology Fellowship Match: How It Failed, and Why It Could Succeed Once Again. *Gastroenterology* **127**:658–666.
Niederle, M. and A. E. Roth. 2009. Market Culture: How Rules Governing Exploding Offers Affect Market Performance. *American Economic Journal: Microeconomics* **1**:199–219.
Niederle, M. and A. E. Roth. 2010. The Effects of a Central Clearinghouse on Job Placement, Wages, and Hiring Practices. In *Labor Market Intermediation*, ed. D. Autor. Chicago: The University of Chicago Press.
Niederle, M. and L. Vesterlund. 2007. Do Women Shy away from Competition? Do Men Compete Too Much? *Quarterly Journal of Economics* **122**:1067–1101.
Ochs, J. and A. E. Roth. 1989. An Experimental Study of Sequential Bargaining. *American Economic Review* **79**:355–384.
Ostrom, E. 1998. A Behavioral Approach to the Rational Choice Theory of Collective Action. *American Political Science Review* **92**:1–22.
Pearson, M. and B. C. Schipper. 2009. The Visible Hand: Finger Ratio (2D:4D) and Competitive Behavior. Unpublished.
Plott, C. R. 1987. Dimensions of Parallelism: Some Policy Applications of Experimental Methods. In *Laboratory Experimentation in Economics: Six Points of View*, ed. A. E. Roth. New York Cambridge University Press.
Plott, C. R. 1997. Laboratory Experimental Testbeds: Application to the PCS Auction. *Journal of Economics and Management Strategy* **6**:605–638.
Rapoport, A., O. Frenkel, and J. Perner. 1977. Experiments with Cooperative 2 × 2 Games. *Theory and Decision* **8**:67–92.
Rassenti, S. J., V. L. Smith, and R. L. Bulfin. 1982. A Combinatorial Auction Mechanism for Airport Time Slot Allocation. *Bell Journal of Economics* **13**:402–417.

Rosser, J. B., Jr. and C. Eckel. 2010. Introduction to JEBO Special Issue on 'Issues in the Methodology of Experimental Economics'. *Journal of Economic Behavior and Organization* **73**:1–2.

Roth, A. E. 1979. *Axiomatic Models of Bargaining, Lecture Notes in Economics and Mathematical Systems #170*. Berlin: Springer Verlag. http://kuznets.fas.harvard.edu/~aroth/Axiomatic_Models_of _Bargaining.pdf

Roth, A. E. 1984. The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. *Journal of Political Economy* **92**:991–1016.

Roth, A. E. 1986. On the Allocation of Residents to Rural Hospitals: A General Property of Two Sided Matching Markets. *Econometrica* **54**:425–427.

Roth, A. E. 1987. Laboratory Experimentation in Economics. *Advances in Economic Theory, Fifth World Congress*, ed. Truman Bewley. New York: Cambridge University Press, pp. 269–299.

Roth, A. E. 1988. Laboratory Experimentation in Economics: A Methodological Overview. *Economic Journal* **98**:974–1031.

Roth, A. E. 1990. New Physicians: A Natural Experiment in Market Organization. *Science* **250**:1524–1528.

Roth, A. E. 1991. A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U.K. *American Economic Review* **81**:415–440.

Roth, A. E. 1993. On the Early History of Experimental Economics. *Journal of the History of Economic Thought* **15**:184–209.

Roth, A. E. 1994. Let's Keep the Con Out of Experimental Econ.: A Methodological Note. *Empirical Economics* **19**:279–289.

Roth, A. E. 1995a. Introduction to Experimental Economics. In *Handbook of Experimental Economics*, eds. J. Kagel and A. E. Roth. Princeton University Press, pp.3–109.

Roth, A. E. 1995b. Bargaining Experiments. In *Handbook of Experimental Economics*, eds. J. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press, pp. 253–348.

RothRoth, A. E. 2002. The Economist as Engineer: Game Theory, Experimental Economics and Computation as Tools of Design Economics. *Econometrica* **70**:1341–1378.

Roth, A. E. 2008a. Deferred Acceptance Algorithms: History, Theory, Practice, and Open Questions. *International Journal of Game Theory* **36**:537–569.

Roth, A. E. 2008b. What Have We Learned from Market Design? *Economic Journal* **118**:285–310.

Roth, A. E. Forthcoming. Market Design. In *Handbook of Experimental Economics,* Volume 2, eds. J. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press.

Roth, A. E. and I. Erev. 1995. Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior* **8**:164–212.

Roth, A. E. and M. K. Malouf. 1979. Game Theoretic Models and the Role of Information in Bargaining. *Psychological Review* **86**:574–594.

Roth, A. E. and J. K. Murnighan. 1982. The Role of Information in Bargaining: An Experimental Study. *Econometrica* **50**:1123–1142.

Roth, A. E. and E. Peranson. 1999. The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *American Economic Review* **89**:748–780.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *American Economic Review* **81**:1068–1095.

Roth, A. E. and F. Schoumaker. 1983. Expectations and Reputations in Bargaining: An Experimental Study *American Economic Review* **73**:362–372.

Roth, A. E. and M. Sotomayor. 1990. Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. *Econometric Society Monograph Series*. New York: Cambridge University Press.

Rubinstein, A. 1982. Perfect Equilibrium in a Bargaining Model. *Econometrica* **50**:97–109.

Rustichini, Aldo. 2009. Is There a Method of Neuroeconomics? *American Economic Journal: Microeconomics* **1**(2):48–59.

Sauermann, H. and R. Selten. 1959. Ein Oligolpolexperiment. *Zeitschrift fur die Gesamte Staatswissenschaft* **115**:427–471.

Schelling, T. C. 1957. Bargaining, Communication, and Limited War. *Journal of Conflict Resolution* **1**:19–36.

Schelling, T. C. 1958. The Strategy of Conflict: Prospectus for a Reorientation of Game Theory. *Journal of Conflict Resolution* **2**:203–264.

Schelling, T. C. 1960. The Strategy of Conflict. Cambridge: Harvard University Press.

Selten, R. and T. Chmura. 2008. Stationary concepts for experimental 2x2 games. *American Economic Review* **98**(3):938–966.

Smith, V. L. 2008. *Rationality in Economics: Constructivist and Ecological Forms*. New York: Cambridge University Press.

Smith, V. L. 2010. Theory and Experiment: What Are the Questions? *Journal of Economic Behavior and Organization* **73**:3–15.

Sobel, Joel. 2009. Neuroeconomics: A Comment on Bernheim. *American Economic Journal: Microeconomics* **1**(2):60–67.

Starmer, C. 1999. Experiments in Economics: Should We Trust the Dismal Scientists in White Coats? *Journal of Economic Methodology* **6**:1–30.

Sugden, R. 2009. Credible Worlds, Capacities and Mechanisms. *Erkenntnis* **70**:3–27.

Thurstone, L. L. 1931. The Indifference Function. *Journal of Social Psychology* **2**:139–167.

Tversky, A. and D. Kahneman. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* **5**:297–323.

Ünver, M. U. 2001. Backward Unraveling over Time: The Evolution of Strategic Behavior in the Entry-Level British Medical Labor Markets. *Journal of Economic Dynamics and Control* **25**:1039–1080.

Ünver, M. U. 2005. On the Survival of Some Unstable Two-Sided Matching Mechanisms. *International Journal of Game Theory* **33**:239–254.

Wallis, W. A. and M. Friedman. 1942. The Empirical Derivation of Indifference Functions. In *Studies in Mathematical Economics and Econometrics*, eds. O. Lange, F. McIntyre, and T. O. Yntema. Chicago: University of Chicago Press, pp. 175–189.

PART II

ECONOMIC THEORY AND EXPERIMENTAL ECONOMICS

CHAPTER 2

THE RELATIONSHIP BETWEEN ECONOMIC THEORY AND EXPERIMENTS

DAVID K. LEVINE AND JIE ZHENG

Introduction

The relationship between economic theory and experimental evidence is controversial. From reading the experimental literature, one could easily get the impression that economic theory has little or no significance for explaining experimental results. The point of this essay is that this is a tremendously misleading impression. Economic theory makes strong predictions about many situations and is generally quite accurate in predicting behavior in the laboratory. In most familiar situations where the theory is thought to fail, the failure is to properly apply the theory and not that the theory failed to explain the evidence.

That said, economic theory still needs to be strengthened to deal with experimental data: The problem is that in too many applications the theory is correct only in the sense that it has little to say about what will happen. Rather than speaking of whether the theory is correct or incorrect, the relevant question turns out to be whether it is useful or not useful. In many instances it is not useful. It may not be able to predict precisely how players will play in unfamiliar situations.[1] It buries too much in individual preferences without attempting to understand how individual

preferences are related to particular environments. This latter failing is especially true when it comes to preferences involving risk and time, as well as preferences involving interpersonal comparisons—altruism, spite, and fairness.

By way of contrast, in many circumstances equilibrium is robust to modest departures from assumptions about selfish and rational behavior. In these circumstances, the simplest form of the theory—Nash equilibrium with selfish preferences—explains the data quite well. In this case, as we shall explain, predictions about aggregate behavior are quite accurate. Predictions about individual behavior are better explained by a perturbed form of Nash equilibrium—now widely known as quantal response equilibrium.

Equilibrium Theory that Works

The central theory of equilibrium in economics is that of Nash equilibrium. Let us see how that theory works in a reasonably complex voting situation. The model is adapted from Palfrey and Rosenthal (1985). There are voters divided into two groups, namely, supporters of candidate A and supporters of candidate B. The number of voters is odd and divisible by three and can take on the values {3, 9, 27, 51}. Unlike the groups used by Palfrey and Rosenthal, the two groups are not equal in size; that is, group B is larger than group A. In the landslide treatment, there are twice as many members of B as of A. In the tossup treatment, there is one more voter in group B than in group A. The voters may either vote for their preferred candidate or abstain, and the rule is simple majority. The members of the winning group receive a common prize of 105, while those in the losing group receive 5. In case of a tie, both groups receive 55. Voting is costly: The costs are private information and are drawn independently and randomly on the interval [0, 55]. Players are told the rules in a common setting, and they get to play 50 times.

Computing the Nash equilibrium of this game is sufficiently difficult that it cannot be done by hand, nor is it possible to prove that there is a unique equilibrium. However, the equilibrium can be computed numerically, and grid searches show that there is only one equilibrium. The key to equilibrium is the probability of pivotal events: The benefit of casting a vote depends on the probability of being pivotal in an election. Thus a good test of Nash equilibrium is to compare the theoretical probability of a voter being pivotal—that is, of a close election—versus the empirical frequency observed in the laboratory. The graph in Figure 2.1 from Levine and Palfrey (2007) plots the theoretical probability on the horizontal axis and the empirical frequency on the vertical axis. If the theory worked perfectly, the points should align on the 45-degree line. They do. Despite the fact that both theoretically and from observing 50 data points it is no easy matter to infer the probability of being pivotal, the theory works nearly perfectly.

It deserves emphasis that when we speak of "theory" here we are speaking entirely of a theoretical computation. In finding the Nash equilibrium probabilities

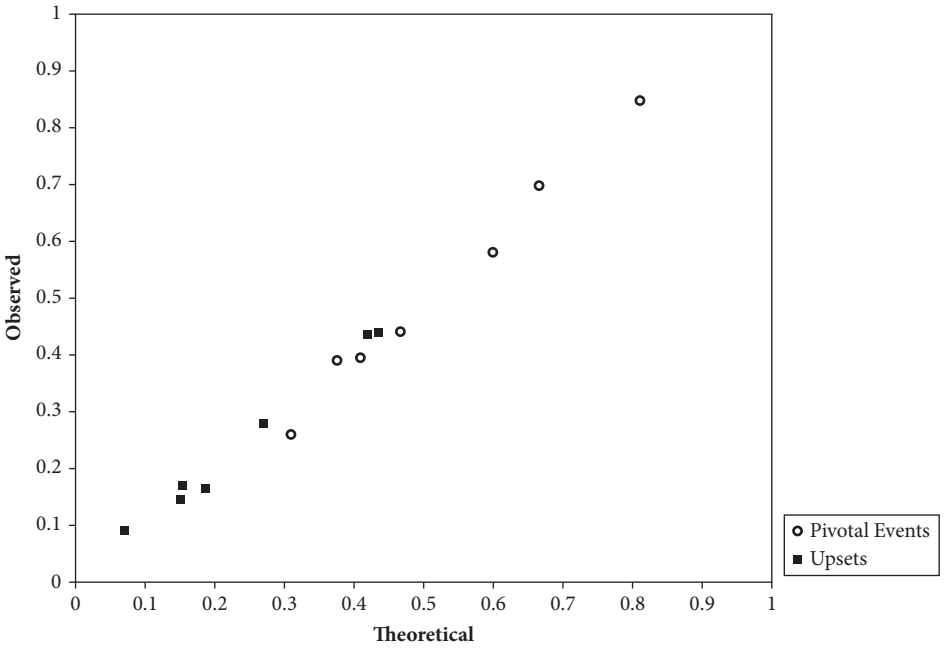


Figure 2.1. The relationship between theoretical probability and empirical frequency for the complex voting example of Levine and Palfrey (2007).

of being pivotal, no parameters are fit to the data: No estimation is done whatever. A pure computation is compared to live data, and the fit is nearly perfect.

The other central theory in economics besides Nash equilibrium is the competitive equilibrium of a market. In modern theory, this can be viewed as the Nash equilibrium of a mechanism in which traders reveal preferences to a market that then determines the equilibrium—with the exact details of the market clearing mechanism of no importance. Experiments on competitive equilibrium—generally in which the market clearing mechanism is a double oral auction in real time—have been conducted many times, dating back at least to the work of Smith (1962). The results are highly robust: Competitive equilibrium predicts the outcome of competitive market experiments with a high degree of accuracy, with experimental markets converging quickly to the competitive price. One typical picture is the history of bids in an experiment by Plott and Smith (1978) showing the convergence to the competitive equilibrium at a price of 60 (Figure 2.2). Again note that the competitive price of 60 is computed from purely theoretical considerations—no parameters are fit to the data.

This picture of data that nearly perfectly fits purely theoretical computations is true for a wide variety of experiments and is very much at odds with the viewpoint that experimental results somehow prove the theory wrong. Indeed the theory fits much better than models that must be estimated in order to fit noisy field data.

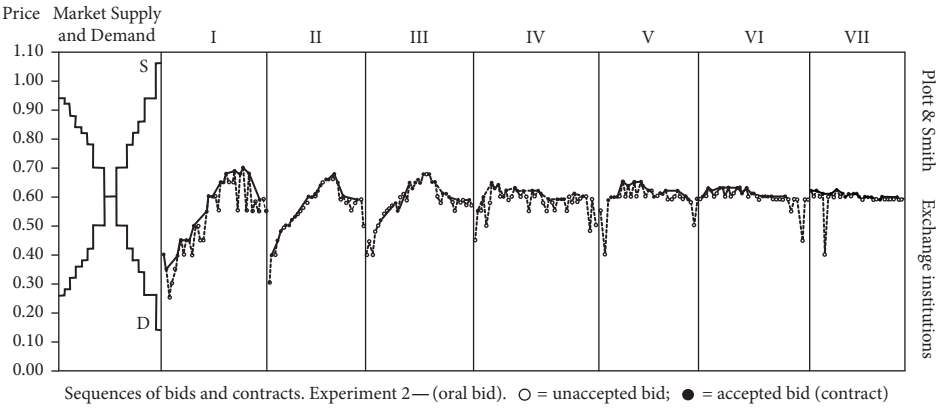


Figure 2.2. The history of bids in experiment 2 (oral bid) by Plott and Smith (1978).

Equilibrium Theory that Does Not Fail

Moving past theory that predicts accurately and well, there are a set of experiments in which equilibrium—especially the refinement of subgame perfection—apparently fails badly. One such example is the ultimatum bargaining game. Here one player proposes a division of $10 in nickels, and the second player may either accept or reject the proposal. If she accepts, then the money is divided as agreed upon. If she rejects, then the game ends and neither player receives any money. Subgame perfection predicts that the second player should accept any positive amount, and so the first mover should get at least $9.95. Table 2.1, with the data from Roth et al. (1991), shows that this is scarcely the case. Nobody offers less than $2.00 and most offers are for $5.00, which is the usual amount that the first player earns. Superficially, it would be hard to imagine a greater rejection of a theory than this. Moreover, like competitive market games, these results have been replicated many times under many conditions.

Despite appearances, theory is consistent with these results—it is the misapplication of the theory that leads to the apparent anomaly. First, the computation of the subgame equilibrium is based on the assumption that players are selfish—that they care only about their own money income. This assumption—which has nothing to do with equilibrium theory, but is merely an assertion about the nature of players' utility functions—is clearly rejected by the data. A selfish player would not reject a positive offer; this fact is the basis for calculating the subgame perfect equilibrium. However, the data clearly show that 5 out of 27 positive offers are rejected. The data—not to speak of common sense—show that many players find low offers offensive in the sense that they prefer nothing at all to a small share of the pie. A "theory" based on the assumption of selfish preferences will naturally

Table 2.1. Offers and Rejection Probabilities for the Ultimatum Bargaining Game

X(\$)	Number of Offers	Rejection Probability (%)
2.00	1	100
3.25	2	50
4.00	7	14
4.25	1	0
4.50	2	100
4.75	1	0
5.00	13	0
	27	

US \$10.00 stakes games, round 10
Source: Roth et al. (1991)

fail to explain the data. However, there is nothing in the logic of rationality, Nash equilibrium, or subgame perfection that requires players to have selfish preferences.

In the mainstream theory of competitive markets, it is true that economists typically assume that people are selfish. This is 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 behavior. Consequently, it is convenient for computational purposes to model people in those environments as being selfish. That should not be taken to mean that this useful modeling tool should be ported to other inappropriate environments, such as bargaining situations.

Surprisingly, even the theory of selfish preferences does not do so badly as a cursory inspection of the data might indicate. Nash equilibrium—as opposed to subgame perfection—allows any offer to be an equilibrium: It is always possible that any lower offer than the one the first player makes might be rejected with probability one, while the current offer is accepted. Nash equilibrium rules out two less obvious features of the data. It rules out a heterogeneity of offers, and it rules out offers being rejected in equilibrium (if players are truly selfish). It is a mistaken view of the theory that leads to the conclusion that this is a large discrepancy. Any theory is an idealization. Players' exact preferences, beliefs, and so forth are never going to be known exactly to the modeler. As a result, the only meaningful theory of Nash equilibrium is Radner's (1980) notion of epsilon equilibrium. This requires only that no player loses more than epsilon compared to the true optimum—which in practice can never be known by the players. The correct test of the goodness of fit of Nash equilibrium in experimental data is not whether the results look like a Nash equilibrium, but rather whether players' losses (epsilon) are small relative to what they might have had.

The correct calculation of the departure of the facts from the theory, in other words, is to (a) determine how much money a player who had available the experimental data could have earned and (b) compare it to how much that player actually earned. To the extent that this is a large amount of money, we conclude that the theory fits poorly. To the extent that it is a small amount of money, we conclude that the theory fits well. This is regardless of whether the data "appear like" a Nash equilibrium or not. The key point is that allowing a small epsilon in certain games can result in a large change in equilibrium behavior. This large change does not contradict the theory of equilibrium—it is predicted by the theory of equilibrium.

For the ultimatum game, Fudenberg and Levine (1997) calculated the losses that players suffered from playing less than optimal strategies given the true strategies of their opponents. Out of the $10 on the table, players only lose on average about $1.00 per game.

This is not the end of the story, however. Nash equilibrium, at least as it is currently viewed, is supposed to be the equilibrium in which players understand their environment, including how their opponents play. It is supposed to be the outcome of a dynamic process of learning—indeed, it may accurately be described as a situation where no further learning is possible. This is important in the games in which the theory worked: In the voting experiment, players played 50 times and thus had a great deal of experience. Similarly, in the double oral auctions, players got to participate in many auctions and equilibrium occurs only after they acquire experience. In the ultimatum game, players got to play only 10 times. More important, in an extensive form game where players are informed only of the outcomes and not their opponents' strategies, players would have to engage in expensive active learning to achieve a Nash equilibrium; and without a great deal of repetition and patience, they have no incentive to do so. In ultimatum bargaining in particular, the first mover can only conjecture what might happen if she demanded more—in 10 plays there is relatively little incentive or opportunity to systematically experiment with different offers to see which will be rejected or accepted. If the game were played 100 times, for example, then it would make sense to try demanding a lot to see if perhaps the opponent would be willing to accept bad offers. In 10 repetitions such a learning strategy does not make sense.

A weaker theory than Nash equilibrium—but one more suitable to the ultimatum bargaining environment—is that of self-confirming equilibrium introduced in Fudenberg and Levine (1993). This asserts that players optimize given correct beliefs about the equilibrium path, but does not require that they know correctly what happens off the equilibrium path, as they do not necessarily observe that. This makes a difference when computing the amount of money players "lose" relative to the true optimum. As we observed in ultimatum bargaining, the first movers cannot know what will happen if they demanded more. So setting a demand that is too low is not a "knowing" error, in the sense that the player has no way to know whether it is an error or not. This leads us to compute not just the losses made by a player relative to the true optimum, but to compute how many of those losses are

"knowing losses," meaning that the player might reasonably know that he is making a loss. Self-confirming equilibrium is a theory that predicts that knowing losses should be low—but makes no prediction about unknowing losses.

For the ultimatum game, Fudenberg and Levine (1997) also calculated the knowing losses. On average, players lose only $0.33 per game, and this is due entirely to second players turning down positive offers—which as we noted has nothing to do with equilibrium theory at all. It is interesting to compare the impact of preferences (the spiteful play of the second players) versus that of learning (the mistaken offers of the first players). On average, players lose $0.33 due to having preferences that are not selfish, and on average they lose $0.67 because they lack adequate opportunity to learn about their opponents' strategies. The losses due to the deviation of preferences from the assumption of selfish behavior are considerably less than the losses due to incomplete learning.

The message here is not that theory does well with ultimatum bargaining. Rather the message is that theory is weak with respect to ultimatum bargaining—very little data in this game could be inconsistent with the theory. Rather, by applying the theory inappropriately, the conclusion was reached that the theory is wrong, while the correct conclusion is that the theory is not useful. Modern efforts in theory are quite rightly directed toward strengthening the theory—primarily by better modeling the endogenous attitudes of players toward one another as in Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), or Gul and Pesendorfer (2004).

We can tell a similar tale of poorly applied subgame perfection in the other famous "rejection" of theory, the centipede game of McKelvey and Palfrey (1992) (Figure 2.3).

The extensive form of the game is shown below. There are two players, and each may take 80% of the pot or pass, with the pot doubling at each round. Backwards induction says to drop out immediately. In fact, as the empirical frequencies in the diagram show, only 8% of players actually do that. As in ultimatum bargaining, the evidence seems to fly in the face of the theory. Again, a closer examination shows that this is not the case.

In a sense, this centipede game is the opposite of ultimatum. In ultimatum the apparent discrepancy with theory was driven by the fact that second movers are

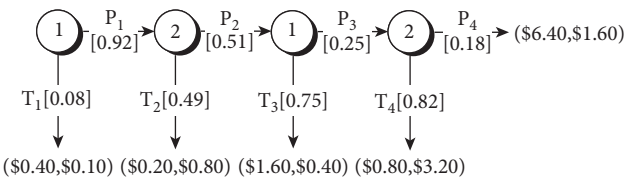


Figure 2.3. The centipede game of McKelvey and Palfrey (1992): Numbers in square brackets correspond to the observed conditional probabilities of play at each information set in rounds 6–10.

spiteful in the sense of being willing to take a small loss to punish an ungenerous opponent. In centipede the discrepancy is driven by altruism—by the willingness of a few players to suffer a small loss to provide a substantial reward to a generous opponent. The crucial empirical fact is that 18% of players will make a gift to their opponent in the final round. Note that it costs them only $1.60 to give a gift worth $5.60. These gifts change the strategic nature of the game completely. With the presence of gift-givers, the true optimal strategy for each player is to stay in as long as possible. If you are the first mover, stay in and hope you get lucky in the final round. If you are the second mover and make it to the final round, go ahead and grab then.

Most of the losses in centipede are actually suffered by players (foolishly misapplying subgame perfection?) who do not realize that they should stay in as long as possible, and so they drop out too soon. Overall losses were computed by Fudenberg and Levine (1997) to be about $0.15 per player per game. However, if you drop out too soon, you never discover that there were players giving money away at the end of the game, so those losses are not knowing losses. The only knowing losses are the gifts by players in the final round. These amount to only $0.02 per player per game. Note that as in ultimatum, failed learning is responsible for substantially greater losses than deviation in preferences from the benchmark case of selfishness.

Another important effort is to try to capture the insight of epsilon equilibrium—that when some players deviate a little from equilibrium play, this may greatly change the incentives of other players—without losing the predictive power of Nash equilibrium. The most important effort in that direction is what has become known from the work of McKelvey and Palfrey (1995) as quantal response equilibrium. This allows for the explicit possibility that players make random errors. Specifically, if we denote the utility that a player receives from her own pure strategy s_i and opponents mixed strategy σ_{-i} by $u_i(s_i, \sigma_{-i})$ and let $\lambda_i > 0$ be a behavioral parameter, we define the propensity with which different strategies are played by

$$p_i(s_i) = \exp(\lambda_i u_i(s_i, \sigma_{-i})).$$

Quantal response theory then predicts that the mixed strategies that will be employed are given by normalizing the propensities to add up to 1:

$$\sigma_i(s_i) = p_i(s_i) / \sum_{s_i'} p_i(s_i').$$

This theory, like Nash equilibrium, makes strong predictions. As $\lambda_i \longrightarrow \infty$, these predictions in fact converge to those of Nash equilibrium. One important strength of this theory is that it allows for substantial heterogeneity at the individual level. This is important, because experimental data are quite noisy and individual behavior is generally heterogeneous.

A good example of this is in the Levine and Palfrey (2007) voting experiment described in the first section. The aggregate fit of the theory was very good; but

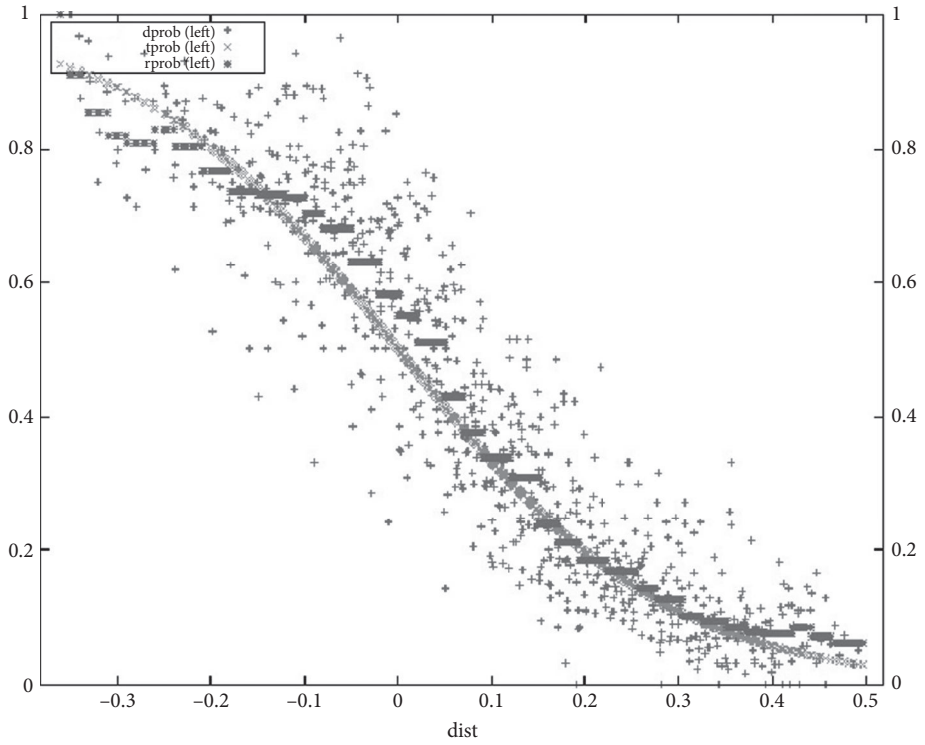


Figure 2.4. The relationship between turnout rate and the loss from participating for the complex voting example of Levine and Palfrey (2007).

at the individual level, the theory fits poorly. Figure 2.4 taken from that paper shows the empirical probability with which a voter participates as a function of the loss from participating. If the loss is positive, Nash equilibrium predicts that the probability of participation should be zero; if it is negative, the probability of participation should be one, and the data should align themselves accordingly. The individual data, represented by plus-signs and the aggregated data represented by the darker lines show that this is by no means true. When losses and gains are small, the probability of participation is relatively random—near 50%. As the loss from participating increases, the probability of participating decreases— but it hardly jumps from 1 to 0 as the threshold of indifference is crossed. However, the gradual decline seen in the data is exactly what is predicted by quantal response equilibrium. Quantal response predicts that when players are near indifferent, they effectively randomize. As incentives become stronger, they play more optimally. The downward sloping curve shows the best-fit quantal response function, where λ_i is estimated from the data. As can be seen, it fits the individual level data quite well.

A key idea here is that in the aggregate, quantal response equilibrium may or may not be sensitive to values of λ_i that are only moderately large. In some games,

such as the voting game, it makes little difference to aggregate behavior what λ_i is, since some voters over-voting makes it optimal for other voters to under-vote. Similarly in the market games, individual errors do not matter much at the aggregate level. The important thing is that we can always compute the quantal response equilibrium and determine how sensitive the equilibrium is to changes in λ_i.

A good illustration of the strength—and potential weakness—of quantal response equilibrium is the mixed strategy example of the asymmetric matching pennies game, described in Goeree and Holt (2001).[2] It is a simple simultaneous game where the row player chooses between *Top* and *Bottom* and the column player chooses between *Left* and *Right*. The payoff is (40, 80) when the outcome is (*Top*, *Right*) or (*Bottom*, *Left*), and it is (80, 40) when the outcome is (*Bottom*, *Right*). It would be symmetric if the payoff for the outcome (*Top*, *Left*) was (80, 40), but here we are interested in the asymmetric cases where the payoff for (*Top*, *Left*) is (320, 40) in one case (denoted "the (320, 40) case") and (44, 40) in the other (denoted "the (44, 40) case"). The data in Goeree and Holt (2001) show in the lab that 96% of row players play *Top* and that 16% of column players play *Left* in the (320, 40) case, with the fraction numbers 8% and 80% respectively for the (44, 40) case. It is obvious that these lab results are quite different from what the theory of Nash equilibrium predicts, where the fraction of row players playing *Top* should be 50% in both cases.

If we apply the theory of quantal response equilibrium to this mixed strategy example, the prediction power can be improved by a large degree. We do the calculations using each of the two alternative assumptions[3]: (1) the standard selfish preference assumption ($U_i = u_i$) and (2) the more realistic altruistic preference assumption ($U_i = \alpha u_i + (1-\alpha)u_{-i}$, where $\alpha \in [0, 1)$). In Figure 2.5, the horizontal axis represents the fraction of row players who play *Top* and the vertical axis represents the fraction of column players who play *Left*. Both Nash and quantal response equilibria are shown: The original equilibrium corresponding to the selfish case and the "new" equilibrium corresponding to altruistic preference with parameter $\alpha = 0.91$ are shown. The curves correspond to different quantal response equilibria with different values of λ. Note that we assume $\lambda_1 = \lambda_2 = \lambda$ since players are drawn from the same population.

By allowing players to make mistakes, as we can see from the graph, the theory of quantal response equilibrium gives a better prediction than Nash equilibrium does. This is especially true in the (320, 40) case with altruistic preference assumption: When $\lambda = 20$, the quantal response equilibrium is quite close to what the experimental data show. It is also worth noting from the graph that the improvement in results from applying quantal response equilibrium alone (for example, in the (320, 40) case, equilibrium shifted from (0.5, 0.13) to (0.82, 0.22)) is more than the improvement from assuming altruistic preference alone (respectively, equilibrium shifted from (0.50, 0.13) to (0.75, 0.12)). What remains mysterious is the (44, 40) case, where the lab result is poorly explained either by allowing people to make mistakes or by the preference of altruism.

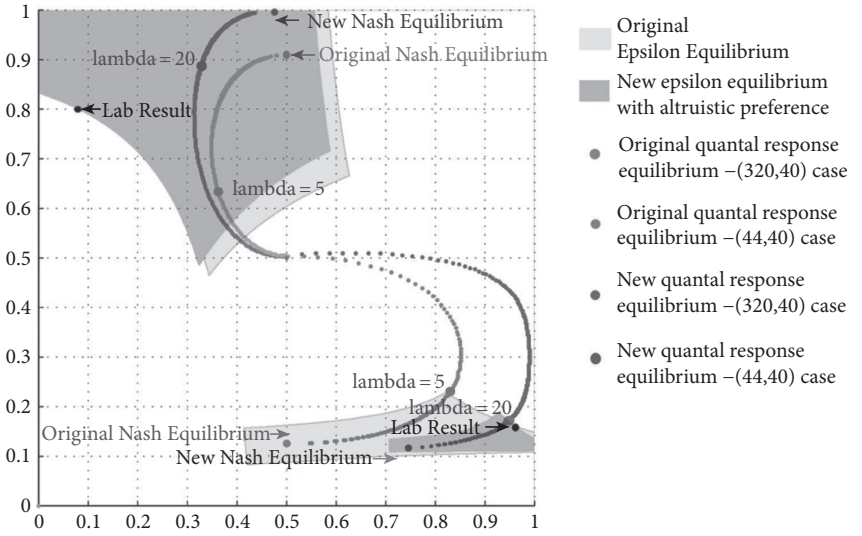


Figure 2.5. Quantal response equilibrium and ε equilibrium for the asymmetric matching pennies game of Goeree and Holt (2001).

We can also analyze ε equilibrium in this game. Under the selfish preference assumption, the laboratory data correspond to an value of \$0.07 per player per game for the (44, 40) case and \$0.06 for the (320, 40) case. The maximum possible amount that could be earned in each game is \$0.80 for the (44, 40) case and \$3.20 for the (320, 40) case. The ε values are \$0.05 and \$0.02 respectively under the altruistic preference assumption. In the (44, 40) case the set of possible equilibria is quite large: Pretty much any mixture in which the fraction of row players playing *Top* is less than 50% and the fraction of column players who play *Left* is greater than 50% is an ε equilibrium. In this sense it is not surprising that the lab result is far off the prediction of the selfish-rational theory. What is interesting is that the perturbations to payoffs that explain the laboratory result are neither due to errors (quantal response) nor due to altruism.

In the (320, 40) case, the set of equilibria is not so large. It predicts little about the row players' play—just that the row player should play *Top* more than 50% of the time. This must be the case, as the Nash equilibrium requires 50% play *Top*, and in the laboratory result 96% play *Top*, and of course both of these must lie in the equilibrium set. Note, however, that the play of the column player is predicted with a relatively high degree of precision: It lies in the range of 10–22%.

What Experiments Have Taught Us

Experimental economics has certainly taught us where the theory needs strengthening—as well as settling some long-standing methodological issues. For

example, the issue of "why should we expect Nash equilibrium" has always had two answers. One answer is that players introspectively imagine that they are in the shoes of the other player, and they reason their way to Nash equilibrium. This theory has conceptual problems, especially when there are multiple equilibria. It also has computational issues—for example, there is a great deal of evidence that the game in which commuters choose routes to work during rush hour is in equilibrium, although individual commuters certainly do not compute solutions to the game. Nevertheless, in principle, players might, at least in simpler games, employ a procedure such as the Harsanyi and Selten (1988) tracing procedure. Experimental evidence, however, decisively rejects the hypothesis that the first-time players are exposed to a game they manage to play a Nash equilibrium. As a result, the current view—for example, in Fudenberg and Levine (1998)—is that if equilibrium is reached, it is through learning. For example, the rush hour traffic game is known from the work of Monderer and Shapley (1996) to be a potential game, and such games have been shown, for example by Sandholm (2001), to be stable under a wide variety of learning procedures.

As Nash equilibrium cannot predict the outcome of one-off games, one area of theoretical research is to investigate models that can. The most promising models are the type models of Stahl and Wilson (1995): Here players are viewed as having different levels of strategic sophistication. At the bottom level, players play randomly; more sophisticated players optimize against random opponents; players who are even more sophisticated optimize against opponents who optimize against random opponents, and so forth. Experimental research, for example by Costa-Gomes et al. (2001), shows that these models can explain a great deal of first-time play, as well as the details of how players reason. The greatest lacuna in this literature is that it has not yet been well tied in to a theory of learning: we have a reasonable theory of first-time play and a reasonable theory of long-term play, but the in-between has not been solidly modeled.

The second area we highlighted above is the area of interpersonal preferences: altruism and spite. As mentioned, there are a variety of models including Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), or Gul and Pesendorfer (2004), that attack this problem, but there is not as yet a settled theory.

There is one "emperor has no clothes" aspect of experimental research. This involves attitudes toward risk. The standard model of game theory supposes that players' preferences can be represented by a cardinal utility function. The deficiency in this theory was highlighted by Rabin's (2000) paradox

> Suppose we knew a risk-averse person turns down 50–50 lose $100/gain $105 bets for any lifetime wealth level less than $350,000, but knew nothing about the degree of her risk aversion for wealth levels above $350,000. Then we know that from an initial wealth level of $340,000 the person will turn down a 50–50 bet of losing $4,000 and gaining $635,670.

The point here is that in the laboratory, players routinely turn down 50–50 lose $100/gain $105 gambles and even more favorable gambles. Yet this is not only inconsistent with behavior in the large, it is off by (three!) orders of magnitude. Roughly, the stakes in the laboratory are so small that any reasonable degree of risk aversion implies risk neutrality for laboratory stakes—something strongly contradicted by the available data.

There are various possible theoretical fixes, ranging from the prospect theory of Tversky and Kahneman (1974) to the dual-self approach of Fudenberg and Levine (2006), but it is fair to say that there is no settled theory and that this is an ongoing important area of research.

Conclusion

The idea that experimental economics has somehow overturned years of theoretical research is ludicrous. A good way to wrap up, perhaps, is with the famous prisoner's dilemma game. No game has been so much studied either theoretically or in the laboratory. One might summarize the widespread view as follows: People cooperate in the laboratory when the theory says they should not. *Caveat emptor*. The proper antidote to that view can be found in the careful experiments of Dal Bo (2005). The proper summary of that paper is as follows: Standard Nash equilibrium theory of selfish players works quite well in predicting the laboratory behavior of players in prisoner's dilemma games.

What experimental economics has done very effectively is to highlight where the theory is weak, and there has been an important feedback loop between improving the theory—quantal response equilibrium being an outstanding example—and improving the explanation of experimental facts.

NOTES

We are grateful to NSF grant SES-03-14713 for financial support, to Drew Fudenberg and Tom Palfrey for many conversations on this topic, to Colin Camerer for helpful comments, and to Guillaume Frechette for encouraging us to do this.

1. The theory sometimes can still make a good prediction even when players are not familiar with the game being played. See Camerer (2003) for examples.
2. Note, however, that players only got to play once, so no learning was possible.
3. We also did the calculation by assuming that a fraction $(1 - \beta)$ of people have altruistic preference and that the rest ($\beta \in (0, 1)$) of the people are selfish, but the result is not improved much from the case in which $\beta = 0$, which is equivalent to assumption 2.

REFERENCES

Bolton, G. E. and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* **90**:166–193.

Camerer, C. F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Costa-Gomes, M., V. P. Crawford, and B. Broseta. 2001. Cognition and Behavior in Normal-Form Games: An Experimental Study. *Econometrica* **69**(5):1193–1235.

Dal Bo, P. 2005. Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *American Economic Review* **95**(5):1591–1604.

Fehr, E. and K. M. Schmidt. 1999. A Theory Of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* **114**(3):817–868.

Fudenberg, D. and D. K. Levine. 1993. Self-Confirming Equilibrium. *Econometrica* **61**(3):523–545.

Fudenberg, D. and D. K. Levine. 1997. Measuring Players' Losses in Experimental Games. *Quarterly Journal of Economics* **112**(2):507–536.

Fudenberg, D. and D. K. Levine. 1998. *The Theory of Learning in Games*, MIT Press.

Fudenberg, D. and D. K. Levine. 2006. A Dual Self Model of Impulse Control. *American Economic Review* **96**(5):1449–1476.

Goeree, J. K and C. A. Holt. 2001. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *American Economic Review* **91**(5):1402–1422.

Gul, F. and W. Pesendorfer. 2004. The Canonical Type Space for Interdependent Preferences. Unpublished.

Harsanyi, J. C. and R. Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

Levine, D. K. 1998. Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* **1**(3):593–622.

Levine, D. K. and T. R. Palfrey. 2007. The Paradox of Voter Participation: A Laboratory Study. *American Political Science Review* **101**:143–158.

McKelvey, R. D. and T. R. Palfrey. 1992. An Experimental Study of the Centipede Game. *Econometrica* **60**(4):803–836.

McKelvey, R. D. and T. R. Palfrey. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* **10**:6–38.

Monderer, D. and L. S. Shapley. 1996. Potential Games. *Games and Economic Behavior* **14**:124–143.

Palfrey, T. R. and H. Rosenthal. 1985. Voter Participation and Strategic Uncertainty. *American Political Science Review* **79**(1):62–78.

Plott, C. R. and V. L. Smith. 1978. An Experimental Examination of Two Exchange Institutions. *Review of Economic Studies* **45**:133–153.

Rabin, M. 2000. Risk Aversion and Expected-Utility Theory: A Calibration Theorem. *Econometrica* **68**(5):1281–1292.

Radner, R. 1980. Collusive Behavior in Noncooperative Epsilon-Equilibria of Oligopolies with Long but Finite Lives. *Journal of Economic Theory* **22**(2):136–154.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *American Economic Review* **81**(5):1068–1095.

Sandholm, W. H. 2001. Potential Games with Continuous Player Sets. *Journal of Economic Theory* **91**:81–108.

Smith, V. L. 1962. An Experimental Study of Competitive Market Behavior. *Journal of Political Economy* **70**(2):111–137.

Stahl, D. O. and P. W. Wilson. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior* **10**:218–254.

Tversky, A. and D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* **185**:1124–1131.

CHAPTER 3

ON THE RELATIONSHIP BETWEEN ECONOMIC THEORY AND EXPERIMENTS

ANDREW SCHOTTER

Introduction

Let me begin by relating a story. I was recently a member of my university's promotion and tenure committee, and one of the cases we discussed was that of a young scientist being brought up for tenure by his department. When I read the docket, I was stuck by a sentence in one of the candidate's letters, which said, in essence, that if this theorist's work was supported by upcoming experiments, his work would have to be considered a major accomplishment.

What struck me about this statement was not what it said about the natural sciences, but rather what it implied about economics, and more specifically economic theory. In all of my years of hanging out with economic theorists, I do not think I ever heard one of them say, "I have just written a theoretical paper, and I hope that, when it is tested, it will prove correct, since otherwise all my work will have been for naught." On the contrary, my experience with theorists is one where they say, "I have just written a theoretical paper; is it not beautiful and elegant?" Historically, theorists in economics have not looked to experimentalists for validation, because they either feel that testing is irrelevant or believe that experiments cannot be trusted to reveal the truth. (Read your referee reports. When experiments fail to

corroborate a theory in the lab, it somehow always seems to be a failure of the experiment and not the theory.) Most economic theorists, like mathematicians, judge theories by their elegance, not by their potential for empirical validation.

This attitude is very different from that found in physics. Take Einstein and his theory of general relativity, for example. In the history of that theory, experimental results were intimately tied to the validation of the theory and its legitimacy. In fact, in 1920 Einstein is quoted as saying:

> If the observations of the red shift in the spectra of massive stars don't come out quantitatively in accordance with the principles of general relativity, then my theory will be dust and ashes. Feyerabend (1993, p. 42, footnote 12).

Now, maybe I have been hanging out with the wrong crowd of theorists, but I challenge anyone to offer an equivalent quote from an eminent economic theorist indicating a similar awareness that theory is merely fancy speculation until supported by some evidence.

It is unclear how economic theory has been able to grow and flourish with such limited respect for empirical validation. One possibility is that historically, economic theorists were philosophers whose theories were no more than speculations about human nature; when these speculations were eventually mathematized, the ethos of the field was already set in cement. Another possibility is that when economic theory was first created and later formalized in the twentieth century, it was strongly believed that economics was not an experimental science, so seeking validation through experiments was pointless. Whatever the reason, theorists and experimentalists have historically had an uneasy relationship in economics, at least compared to that in the natural sciences. This is not to say that the theory–experiment relationship in the natural sciences is utopian; from what I have learned, for example, experimentalists tend to get all the funding while theorists get all the prestige.

In this chapter I want to assess the relationship between economic theory and experiments; I want to provide some insight into the nature of that relationship; and I also want to state what it should or must be, given the opportunities and constraints we face as social scientists. While the history of this relationship is one of estrangement and resistance, the future looks different. That future will redound to the mutual benefit of theorists and experimentalists, however, only if they recognize the limitations of testing social–scientific theories in the lab, and specifically the limits on the types of questions that experimentalists are capable of answering. To that end, I want to discuss five questions, all of which deal with the relationship between theory and experiment and the proper way to test theory in the lab. The questions asked are:

1. Why is theory needed for experimentation?
2. What is the difference between predictive and explanatory theory? Is one more valuable than the other?
3. What does it mean to "test theory" in the lab?

4. What level of aggregation should we use to test theory?
5. What features of data present the proper test of the theory?

I offer neither perfect solutions nor great insights. What I am mostly aiming to do is to share my confusion about the relationship between economic theory and experiments and raise a number of issues that I find interesting and puzzling.

What Is the Relationship Between Theory and Experiments in Economics?

Strong and Wrong Theories: The Role of Theory in Experiments

Economics is known as the queen of the social sciences because of the rigor it brings to answering questions posed to it. What this rigor means is that economists look to understand the exact circumstances under which various statements are correct. For example, take the following statement:

Statement 1. When the price of a good increases, demand falls.

This statement, while passing the test of common sense, is not a correct statement. Whether the demand for a good falls when the price increases depends on the relationship between the income and substitution effect at the current price—that is, whether the good is a Giffen good or not. So what economic theory does is provide the conditions under which Statement 1 is correct. More generally, economic theory is a search for the circumstances that must exist in order for a statement about economic or social behavior to be logically correct. A successful theorist provides all the ifs that are necessary for the logical completion of if–then statements.

By this definition, economic theory has no connection to the real world. Theorists are focused on finding the minimal conditions for statements to be logically correct, and whether those conditions are realistic or descriptive of the real world is a question that is rarely asked. To my mind, however, such theories are wrong by construction. It is just obvious a priori that deducing the minimal conditions for certain statements to be logically correct cannot result in accurate predictions of human behavior. In my view, *all theory is wrong until tentatively validated.*

So how can a science based on a wrong theory function? Why should theory be used to guide experiments? I argue in Schotter (2007) that while economic theory may be "wrong" by definition, it is also "strong" in the sense that its assumptions are transparent, which allows its predictions to be evaluated logically. It is also strong in the sense that it allows economists to make point and interval predictions about

social situations that are not possible for other social sciences whose theories are not as formal and whose assumptions not as transparent.

I will now outline three examples of the usefulness of "strong and wrong" theories: the Ultimatum Game, Expected Utility Theory, and Quantal Response Equilibrium.

The Ultimatum Game

If one adheres to the premises of selfishness and subgame perfection, then economic theory can predict what will happen in the Ultimatum Game. The Proposer will offer zero or epsilon to the Respondent, and the Respondent will accept.[1] No other social science can make these types of predictions. Economics is what I call a "strong" social science because its theories lend themselves to prediction. It is this strength of economic theory that sets it apart from our sister social sciences. In the case of the Ultimatum Game, though, economic theory's predictions are wrong. In my view, strong and wrong theories are useful because they lend structure to the scientific enterprise—specifically, through falsification of old theories and construction of new and better ones.

The history of Ultimatum Game experiments demonstrates this point. Güth (Güth et al., 1982), the Ultimatum Game's experimental pioneer, was interested in whether subjects were capable of performing backward induction as predicted by economic theory. While the subjects proved capable of backward induction, Güth discovered that proposers nevertheless made positive offers.[2] These "anomalies" — and without theory there are no anomalies—were generally thought to result from subjects' violation of either the selfishness axiom or subgame perfection. More recently, the Fehr–Schmidt (Fehr and Schmidt, 1999) and Bolton–Ockenfels (Bolton and Ockenfels, 2000) theories of inequity aversion have provided even better explanations for these anomalies. The economic theory of the Ultimatum Game was wrong, but its transparent assumptions allowed us to make predictions, which paved the way for systematic investigations of how and why it was wrong.[3]

Expected Utility Theory

Expected Utility Theory also demonstrates the usefulness of a strong and wrong theory. Von Neumann and Morgenstern (1944) identified a set of axioms that, if descriptive of the way people think, predicted that they would behave according to the precepts of Expected Utility Theory. Whether these axioms accurately described the world was an empirical question that could be investigated experimentally. Allais (1953), Ellsberg (1961), and others showed that some of the axioms of the theory, most notably the Independence Axiom, did not accurately describe how humans make decisions. Research into these anomalies led to Kahneman and Tversky's (1979) Prospect Theory and a host of other new alternatives to Expected Utility Theory. Without Expected Utility Theory functioning as a straw man, Prospect Theory and other non-expected-utility theories would not have

been developed. It is precisely because Expected Utility Theory was wrong or at least was capable of being violated and these violations proved to be informative in interesting ways.

Quantal Response Theory

Another example of a wrong theory leading to improved theorizing is the case of Quantal Response Equilibrium (QRE), which was designed to remedy the inaccuracies of Nash theory's comparative-static and point predictions. Here, the Nash theory failed to account for the fact that subjects don't behave in accordance with best-response theory; humans make decisions stochastically, with noise, and they take into account the payoffs of decisions relative to other available decisions. QRE accounts for these decision elements, and it results in better qualitative and quantitative predictions. As with the Ultimatum Game and Expected Utility Theory, QRE would likely not have been constructed except for the existence of Nash theory's systematically false predictions.

To summarize, what I have said above should be taken as a justification for theory-led experiments. In my experience, the most effective way to structure experimental research in the social sciences is to use theory as a benchmark (straw man, if you prefer) for the judgment of experimental data. Analysis of deviations from the theory's predictions can help sort out what assumptions are violated. On this view, rather than providing valid explanations of the world, theories codify logically consistent and intelligible assumptions that produce testable predictions. While I certainly root for my own theories and perhaps those of my close friends in the lab, I am equally happy to find interesting deviations as I am to find confirmation. Testing a theory is valuable because it can demonstrate how the theory fails while generating inferences about what elements of human behavior were missing from the theory. Indeed, I would submit that economics has made more progress by analyzing the failures of theory rather than successes.

What Is the Difference Between Predictive and Explanatory Theory? Is One More Valuable than the Other?

What should we expect from good theories? According to Milton Friedman, we should expect them to make good predictions:

> The ultimate goal of a positive science is the development of a "theory" or "hypothesis" that yields valid and meaningful (i.e. not truistic) predictions about phenomena not yet observed. Friedman (1953, p. 7)

Friedman emphasizes that our theories should generate ex ante predictions of events "not yet observed" rather than ex post rationalizations of previously observed phenomena. By "not yet observed," Friedman does not mean events strictly in the future, however:

> [P]redictions by which the validity of a hypothesis is tested need not be about phenomena that have not yet occurred, that is, need not be forecasts of future events, they may be about phenomena that have occurred but observations on which have yet to be made or are not known to the person making the prediction. Friedman (1953, p. 7)

In other words, Friedman's predictiveness also includes "postdictiveness," inasmuch as the past observations consist of untested data.

Because predictiveness is paramount, Friedman asserts that the empirical realism of a theory's assumptions is not a basis for critical evaluation. I reject that view by the following reasoning. Consider Statement X, which follows from assumptions a, b, and c. If Statement X predicts well when tested using real-world data, yet any of the assumptions are wrong, then the theory faces one of three problems: (1) One or more assumptions is superfluous, (2) two or more assumptions counteract each other in a way that is consistent with the data, or (3) the consistency of the data with the theory is altogether spurious. Conversely, if all the assumptions of a model are correct and the model is complete, then the predictions of the theory must be corroborated. It is difficult, therefore, to see how the assumptions of a model are unimportant—after all, the assumptions are the theory.

Many people have argued, and I think rightly, that the theoretical exercise advocated by Friedman is not the only one we can engage in. We are often presented with data that contradict our intuitions or our theoretical presuppositions; these data require explanation, which requires the construction of new theory. While the value of such explanatory theory is evident, Friedman distinctly places it below predictive theory. I think it more accurate to say that experimentalists face a trade-off between prediction and explanation. That is, one must find a balance between (1) theories that make precise predictions but may not be able to explain data effectively ex post, and (2) theories that make weak predictions yet can explain data effectively ex post.

Examples of Predictive and Explanatory Theories

Nash Versus Quantal Response

Let me explore this difference in more detail. Take the difference between the Nash theory and the Quantal Response Theory as it is applied to a simple 2×2 game. Consider the game used by Goeree et al. (2002), where they have people play two different 2×2 games presented to them as follows.

In Game 1 (shown in Table 3.1), as the percentages next to each column and row indicate, subjects behave in accordance to the Nash theory, which predicts

Table 3.1. Game 1: Symmetric Matching Pennies

			Player 2	
		Left (48%)		**Right (52%)**
	Top (48%)	80, 40		40, 80
Player 1				
	Bottom (52%)	40, 80		80, 40

Table 3.2. Game 2: Strongly Asymmetric Matching Pennies

			Player 2	
		Left (16%)		**Right (84%)**
	Top (96%)	320, 40		40, 80
Player 1				
	Bottom (4%)	40, 80		80, 40

that each player, the row and the column player, choose each of his/her strategies with equal probability. This result is what Goeree et al. (2002) call a "treasure," a situation where the theory predicts behavior well. Note that for Game 1 payoffs, QRE predicts the same result as Nash—that is that both players will choose their rows or columns with equal probability no matter what their ability to best respond is (i.e., for any value of λ). For Game 1 payoffs, the theories agree and both make precise, unequivocal predictions.

Now consider Game 2 (shown in Table 3.2), which is identical to Game 1 except for a change in the payoff to the row player in the upper left-hand cell.

In this case the Nash theory again makes a precise prediction: The row player will use the same 50–50 mixed strategy, while the column player will now choose Left 10% of the time and Right 90% of the time. As the percentages indicate, though, this prediction is wrong. Goeree et al. (2002) found that row players choose Top 96% of the time and Bottom 4% of the time, while the column player chooses Left 16% of the time and Right 84% of the time. Nash theory's precise point prediction was wrong.

In contrast to Nash, QRE does not make a precise point prediction for Game 2. To make such a prediction, QRE needs to know what values of λ describe the players' behavior. In other words, QRE is primarily an explanatory theory. It can explain data ex post but cannot make precise predictions ex ante unless one commits to a specific value for λ. Fitting the theory to the data ex post differs from having the theory make a prediction ex ante—what Friedman urges us to do.

In one sense, QRE is predictive, however, in that it can be used to make qualitative predictions about the comparative statics of models. For example, in Goeree and Holt's game, QRE's comparative-static predictions differ significantly from the

Nash theory's (when comparing Games 1 and 2). In Game 2 (relative to Game 1), QRE predicts that the row player will increase his use of Top while the Column player will increase his use of Right. Moreover, QRE makes this prediction robustly; no matter what λ is assumed, QRE correctly predicts the direction of change in behavior between Game 1 and Game 2. Nash theory, meanwhile, predicts no change in behavior for the row player. In this circumstance, then, QRE generates a more accurate (albeit less precise) prediction than Nash theory.

Level-k *Theory*

Predictive theories make predictions before experimental data are collected, while explanatory theories can only be defined after data have been collected. One good place to examine this difference is level-*k* theory (Stahl and Wilson, 1995; Camerer et al. 2004; Costa-Gomes et al. 2001). Crawford and Iriberri (2007) attempt to explain the systematic deviations from unique mixed-strategy equilibrium in zero-sum two-person "hide-and-seek" games with non-neutral framing of locations. Exemplary of such a game is one that was run by Rubinstein, Tversky, and Heller (1993, 1996) (henceforth RTH), which instructed the seekers thusly:

> Your opponent has hidden a prize in one of four boxes arranged in a row. The boxes are marked as shown below: A, B, A, A. Your goal is, of course, to find the prize. His goal is that you will not find it. You are allowed to open only one box. Which box are you going to open?

Common sense advises the hider to put the money in the box that was "least likely." But if the players agreed as to which box was the least likely, then it would become the most likely and would therefore be a silly place to hide the money. By intuition, one notices that the "B" location is distinguished by its label and hence focal, whereas the two "end A" locations may also be inherently focal. These intuitions also give the "central A" location its own brand of uniqueness as the "least salient" location. While the Nash theory predicts that the money will be put in any box with a probability of 0.25, in experiments the hiders put the money in the "least salient" box, "central A," 37% of the time, while seekers look there 46% of the time.

Crawford and Iriberri (2007) try to resolve the following two stylized puzzles:

1. Why do hiders allow seekers to find them 32% of the time when they could hold it down to 25% via the equilibrium mixed strategy?
2. Why do seekers choose central A even more often than hiders?

Success in the hide-and-seek game requires that one outguess his/her opponent. Explaining hide-and-seek decision-making is therefore a matter of level-*k* analysis. Making this analysis more complicated and interesting are the asymmetric labeling of choices and the "least-salient" appearance of "central A."

One should note that without data, we cannot even begin to model this situation using a level-*k* analysis. The key to level-*k* modeling is the construction of what a level-0 player is likely to do. This construction can only be performed after

we have the data we are trying to explain since no theory of level-0 agents exists that predicts this behavior. Crawford and Iriberri (2007) perform this level-0 construction by working from two assumptions. First, they assume that no subjects behave as level-0 types. Anchoring of the level-0 type exists only in the minds of higher-level types as the starting point for their strategic thinking. Second, they assume that the level-0 types for hiders and seekers are identical; that is, they both use identical mixed strategies over the four locations, and their mixed strategies favor salient locations. Formally, level-0 hiders and seekers are assumed to choose A, B, A, A with probabilities $p/2,\ q, 1-p-q,\ p/2$, respectively, with $p > \frac{1}{2},\ q > \frac{1}{4}$, and $p + q \leq 1$. Because the "end A" frequencies are almost equal in the data, Crawford and Iriberri set equal their choice probabilities for level-0 (and higher) types, for simplicity.

Note, however, that these assumptions are made only after the authors know the stylized facts they are trying to explain—that is, only after they have seen the data. Just as QRE needs data to define λ, level-k theory needs data to define what level-0 types do. Like QRE, level-k is an explanatory theory because, given the model's parameters, it cannot be specified until the data it is being fit to are provided. Notwithstanding their weak predictiveness, such exercises are valuable. Crawford and Iriberri's analysis provides an elegant application of level-k theory to a puzzling problem. It is not, however, the type of predictive analysis that Friedman had in mind.

Making Explanatory Theories Predictive

The gap between explanatory and predictive theory is not as wide as the above discussion suggests. Many explanatory theories can use models estimated for previous data sets to make predictions out of sample on other data sets. Crawford and Iriberri, for example, estimate their level-k theory separately for each of RTH's six treatments and use the re-estimated models to "predict" the choice frequencies of the other treatments with a fair amount of success.

More ambitiously, Crawford and Iriberri also test their model's predictive power for data generated by other experiments run independently. They consider two similar experiments, O'Neill's (1987) card-matching game and Rapoport and Boebel's (1992) closely related game. These games both raise the same kinds of strategic issues as RTH's games, but with more complex patterns of wins and losses, different framing, and in the latter case five locations. Using the model estimated from RTH's data, Crawford and Iriberri "predict" subjects' initial responses in these games. Similar exercises are performed by Ho et al. (2008), when they predict their EWA learning model out of sample to demonstrate that they have not over-fit their model to the data.

To summarize, in our everyday lives as a scientists we are faced with two types of theories, namely, those that are predictive and those that are explanatory. While I am not placing one above the other, I do think that it is important to know which type of theory we are working with and not to confuse ex ante prediction with ex post explanation. Theories that explain well may fail to make tight predictions,

while those that make precise predictions may explain data poorly. Friedman would urge us to focus on predictive theories, but that does not mean that explanatory theories are useless. As I have argued and will argue later, if explanatory theories like QRE can be used to make valid comparative-static predictions, then these predictions may be as good as we can expect from social-scientific theories. I have also made the case, however, that predictive theories with poor track records can be useful in that they structure our thinking and provide fodder for later explanatory theories. As we have seen, the failures of the Nash theory led to QRE, just as the failures of Expected Utility Theory lead to Prospect Theory and other theories of decision-making under uncertainty. That is the value of being strong and wrong.

What Does Testing a Theory in the Lab Mean?

I have not argued for a refutation of theory; on the contrary, theory is essential. The real question is *how* we should test our theories or models, not whether models are useful or not. There are two basic approaches to theory-led experimental design and analysis; I call them the structural approach and the comparative-static approach. In this section, I examine these two approaches.

Let us, for the moment, restrict our attention to predictive theories and ask two questions: (1) What is it about the theory that we expect to be testable in the lab? and (2) How should we go about testing it? The structural approach takes a model to the lab in an unaltered state and uses experimental data to estimate the model's parameters by a maximum-likelihood or other goodness-of-fit criterion. The comparative-static approach places less demand on the model and attempts to verify whether, given the parameter values induced in the lab, the predictions of the theory are substantiated. In the strong form of the comparative-static approach, we expect the point predictions of the theory and its comparative statics to be substantiated while, in the weaker form, we expect only the qualitative directional predictions to be supported. A third "quasi-structural" approach assumes a heterogeneous population of behavioral types and uses data to estimate a model of distribution over types. Because the quasi-structural modeler has complete freedom to define the types (including junk categories that soak up all behavior), such models tend to be more ad hoc; I therefore direct my attention in this section to the structural and comparative-static approaches.

The Structural Approach

The structural approach in laboratory experiments borrows directly from the structural approach taken by applied econometricians using field data. Structural theorists often fail, though, to make the proper modifications for the special circumstances existing in the lab. As mentioned, in the structural approach the

model is sacrosanct and should be applied to data unfiltered. Experiments are used not to challenge the model's axioms, but rather to estimate the model's parameters. These parameter values are meaningful in and of themselves and may be exported out of sample to other similar situations.

As shown above, this approach is fundamentally flawed. If a theory is false by construction, it makes no sense to use it as a maintained hypothesis for empirical analysis. In my view, theories should have some empirical validation before we go about calibrating or estimating their parameters. The structural approach, in contrast, just assumes as a null hypothesis that the theory is correct and only uses empirical data to estimate the theory's parameters. By definition, then, the structural approach cannot test theory.

To my knowledge, none of the natural sciences use theory in this manner. For other sciences, the theory generates hypotheses (predictions) for testing, not for sanctification and estimation of parameters. Bringing presumptively false models to data under the maintained hypothesis that they are true is a questionable, if not useless, exercise. However, structural modeling of explanatory theories is a different matter, and if one assumes the fruitfulness of goodness-of-fit measures, they may be useful (see below).

As discussed in Schotter (2008), the main curse we face as economists is that the real world is not arranged so that we can observe the types of variables we'd like to study. For example, beliefs are important in many economic theories, yet they are unobservable. Other unobservables include costs, reservation wages of unemployed workers, and so on. These are things we'd love to be able to measure directly but cannot. The great accomplishment of applied econometrics has been its ability to identify and measure such unobservables through statistical inference. When making inferences, a structural approach brings a theoretical equilibrium apparatus to bear for the analysis of real-world data rather than a partial reduced-form set of equations. This is a "strong and wrong" approach that makes perfect sense given the constraints on the data.

In the lab, however, we are capable of measuring and controlling many of the variables that are unobservable in the real world. Doing experiments therefore allows us to cut down dramatically on inference, and in some lucky cases we can even avoid it altogether by a design that makes all relevant variables directly observable. Such designs normally involve clever treatments that isolate the variables of interest. The upshot is that experimentalists should ask themselves why they would import techniques designed for a world of information scarcity when the entire raison d'être of the lab is the ability to control and measure what in the real world would be unobservable. Why infer when one can observe?

The Comparative-Static Approach

In the comparative-static approach, a theory's predictions are put to the test. The model's comparative-static properties are tested and explored using controlled

treatments. Here there is little attempt to estimate the parameters of the model since they are induced in the design. Instead, great attention is paid to the quantitative and qualitative predictions of the theory. The point predictions of the theory are certainly tested, but experiments are primarily designed to perform the weaker test of whether subject behavior responds in the direction predicted by the theory to changes in the model's parameters.

In my view, the comparative-static approach makes more sense than the structural approach. I am also confident that, for the foreseeable future, it will be more productive scientifically. Given the constraints that we face as social scientists, the comparative-static approach is just about as far as we can go in testing theory. Searching for parameter values that are supposed to be meaningful outside the very narrow confines of a particular experiment is an exercise whose value needs to be thought out carefully.

A Formalization of the Difference[4]

As stated above, one of the problems with the structural approach as practiced by experimentalists is a lack of appreciation for the ability of the lab to allow the investigator to control variables directly. This control obviates the complicated inference strategies necessary when working with field data. Let us contrast the way an econometrician and an experimentalist look at their scientific tasks.

Let v be a vector characterizing an economic agent's (experimental subject's) environment. Elements of v can include probability distributions defined over prices, incomes, and so on. Assume that v is a subset of the space V. The agent possesses an objective function, information set, computational ability, and so on, that defines a behavior which is a function of the environment v. We denote the dependence by the mapping $b : V \rightarrow B$, where B is the space of behaviors.

The pair $\{b, v\}$ map into an outcome y, which may itself be stochastic. Then let $\Gamma : B \times V \rightarrow Y$. This y may be a variable of direct interest to the agent, such as a payoff, or something of indirect interest. It must only be a function of b and v.

The structural econometrician typically only has access to data with values of y. To uncover v in this case, structural analyses take the following route. By assuming a set of objectives and information sets for the agent(s) and assuming unlimited computational ability on the part of the agent(s), an optimal rule may be solved for by the analyst. Assume that such a rule, parametric with respect to v, exists for all $v \in V$, and denote this rule by $b^*(v; \pi)$, where π denotes parameters characterizing preferences, information sets, and/or the computational ability of the agent.

Define a distance function $D(m, n)$ which has standard properties. Then the econometrician estimates $\theta = (\pi, v)$ as follows:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} D(y, \Gamma(b^*(v; \pi), v)),$$

where Θ denotes the parameter space for θ. *Estimation of θ allows the econometrician to perform comparative-static exercises conditional on the rule b^*.*

The experimental approach typically focuses directly on the agent's behavior. Outcome measures (the y discussed above) may have little role to play in an experimental analysis. Instead, the experimentalist can more or less directly measure behavior as he or she varies the environment. Moreover, the experimentalist, through careful experimental design, can induce a certain set of preferences on the subject as well as directly control the subject's information set and computational resources (that is, all elements of π). Let us imagine that π is then perfectly controlled by the experimentalist. The experiment may then proceed by varying v and measuring the behavior of the subject(s) directly. Imagine an experiment in which π is always fixed at value π^0 and where $v \in \{v_1, ..., v_N\}$. Each trial, characterized by a different v, is associated with an observed behavior b_i, $i = 1, ..., N$. Oftentimes the purpose of experimental analysis is to determine the concordance between the predictions of received theory regarding optimal behavior under π^0 and what is actually observed. Let S denote a distance function. Then the predictions of the theory are consistent with observed behavior if

$$\sum_{i=1}^{N} S(b_i, b^*(v_i, \pi^0))$$

is "small."

These distinctions have significance for experimental design. If the objective of the structural approach is to estimate θ from observations on y, for example, then that exercise would call for a design where the experimenter generates as many environments v as possible to provide enough variance in the right-hand variables for efficient estimation of θ. This design recommends very few observations of behavior b at many different values of v. The best way to do this is to generate the v's to be used in the experiment randomly.

Adherents of the comparative-static approach would proceed differently. Under the null hypothesis that the theory is correct, they would choose a few v's in V where the predicted behavior of subjects is quite different and then generate many observations at those particular pre-chosen v's. The downside of this design is that, as stated above, we can only observe the predictiveness of the theory in a small subset of environments over which it is defined.

So as we see from this characterization, importing a structural econometric approach to experimental data fails to take advantage of all of the benefits that experimental and laboratory control offers the experimentalist. It shifts the focus of inference from y to θ under the maintained hypothesis that behavior is characterized by b^*. The experimentalist need not take this approach since he can verify directly if b^* is being followed and can replace inference with observation generally.

Reconciliation

The discussion up to this point has left unanswered the question of the proper relationship between the structural and comparative-static approaches. How can they

both be used to help us better understand the world we live in? My formal approach implies that the lab is ideal for two enterprises: (1) testing theory and (2) estimating, under controlled circumstances, the relationship $b^*(v;\pi)$ for export outside the lab to supplement structural work in the field. I will illustrate this statement by way of two examples.

First, consider the work done on the structural estimation of auction models using field data. Such models require that one estimate the parameters of the bidders' cost distribution, which is unobservable. Assuming as a null hypothesis a strictly monotonic Nash bid function relating cost to bids, the econometrician can estimate the parameters of the bidders' cost distribution. If people systematically deviate from Nash behavior, however, then estimates made under the null are inaccurate.

The assumption that people bid according to a Nash bid function is a testable hypothesis, and laboratory experiments can therefore be of help in this circumstance. Experimentalists can save the econometricians the problem of working from a false assumption by testing the behavioral hypothesis on which the structural estimation of field data is based. If it passes the experimental test, then the econometrician can go ahead and follow the assumption in his/her field estimation. If experiments show, instead, that bidders systematically deviate from Nash behavior, then the econometrician will know that following the assumption in his/her field estimation would lead to faulty results. Instead, he/she will have to seek a more accurate behavioral assumption, perhaps by constructing a bid function from the laboratory data to be used in his identification exercise in the field. (See Bajari and Hortaçsu (2005) for a nice example of a paper that investigates this question.) To account for worries about external validity, one can perform the experiment on bidding professionals using large stakes.[5]

Another example of this method is found in Brown, Flinn, and Schotter (2011) (henceforth BFS). BFS perform a real-time search experiment in which the objects of interest are the subjects' reservation wages and their time path as search costs accumulate in real time. Since these experiments are performed by replicating a stationary environment, the optimal reservation wage policy for the agent is constant over time. This element reflects studies using field data like that performed in Flinn and Heckman (1982), in which the authors used the constant reservation wage policy as a maintained hypothesis in the estimation of unobservable parameters of the search environment—for example, the arrival rate of wages, the parameters of the distribution of wages generating offers, and so on. If there is time dependence in the reservation wage policy, however, then Flinn and Heckman's estimates would be biased. In that case, better estimates could be derived if one replaced the assumption of a constant reservation policy with a time-dependent reservation policy estimated in the lab.

While such a full re-estimation is not attempted by BFS, they do use lab results to estimate the shape of the hazard functions for exits out of unemployment in the 1997 National Longitudinal Survey of Youth. What they find is

that, inconsistent with findings derived from the constant-reservation-policy assumption, the hazard functions based on lab results exhibit both increasing and decreasing segments. Increasing hazards are inconsistent with a constant reservation wage policy for searchers in heterogeneous environments, so BFS's findings indicate time dependence for some agents.

This finding would not have been possible without lab experiments because it is only in the lab that one could exert enough control over the environment facing job searchers to establish the their reservation wages are time-dependent. The declining reservation wage observed in the lab should therefore be incorporated into future work using field data. In this and many other economic investigations, the lab is a place to test the assumptions upon which field work is based. Instead of assuming that bidders use a Nash bid function or that workers search using a fixed reservation wage, we should subject these assumptions to empirical test.

What Can We Learn from Experiments?

Finally, one must ask what to expect to learn from an experimental paper. What is its output? Papers that take the comparative-static approach should test the workings of the theory. We can observe how a theory performs over a carefully selected set of points in the model's parameter space, and we can observe how behavior changes as we move across treatments (across the parameter space). The comparative-static exercise furnishes us with a natural metric by which to judge the performance of the theory. With a good experiment, one can often eyeball the data to see if the theory is doing well. If the theory fails, a good experiment will identify a set of problems with the theory that can lead to new insights and new theories.

A paper that takes the structural approach outputs a goodness-of-fit measure (a likelihood) that shows how close a model, assumed to be true, fits the experimental data. The structural approach does not test the theory, since the truth of the theory is a maintained hypothesis, and it offers no information about what the theory does not explain (Roberts and Pashler, 2000). That is, theories are only useful if there are observations inconsistent with the theory; if the theory can explain any set of observations, it is vacuous.

A structural analysis results in a goodness-of-fit measure whose absolute value is difficult to interpret without a benchmark comparison. In those papers that consider alternative models and likelihoods, the alternatives are rarely as well-crafted or thought-out as that which motivated the paper. They are usually thrown in as straw men for meaningless comparisons. Moreover, the structural approach assumes away what it calls "disequilibrium" behavior because it is impossible for such behavior to exist. In consequence, structural exercises rarely generate unexpected or salient results, since they are ruled out at the outset.

While this may sound critical, it is only an obvious criticism of the structural approach to testing theory in the lab. The main focus of the structural approach in

applied work using field data is the performance of policy thought experiments using equilibrium models with estimated parameters. In other words, the main justification for doing structural work has never been to test theory but rather to examine policy.

Testing Theory

Once one has decided on a theory to take to the lab, a few questions arise. One concerns what level of aggregation you are going to use for your test, and the second involves what characteristics of the data you will want to explain. Let us discuss these one at a time.

What Level of Aggregation Should We Use to Test Theory?

Most microeconomic models brought to the lab are written on the level of the individual. They focus on the choice behavior of the individual agent, and theoretical predictions are made for that individual. What we observe in many experimental papers is that while the theory is one of individual maximization, the data used to test the theory are aggregated over time or across individuals. This approach is usually taken because individual behavior is erratic and heterogeneous, so some smoothing is needed in order to see patterns. Aggregation provides this smoothing. While this disconnect is often harmless, in some cases it gives a misleading account of subject behavior. Let me illustrate this point by reference to Mueller and Schotter (2010), which concerns a test of a model by Moldovanu and Sela (2001) (henceforth M-S) about the proper design of tournaments or contests.

M-S derive the "optimal" set of prizes for an organization trying to motivate workers through an effort contest or tournament. They investigate firms where workers are assumed to be risk-neutral expected-utility maximizers with either linear, convex, or concave cost-of-effort functions and where an organizational designer has a limited amount of money available for bonuses to be awarded to those workers whose outputs are highest. (Assume that output is linear in effort and nonstochastic; effort is essentially equivalent to output, and both are observable.) Workers differ according to their ability, which is randomly assigned from a distribution that is common knowledge.

M-S demonstrate that for organizations where workers have linear or concave cost-of-effort functions, the optimal prize structure is one where the entire prize budget is allocated to one big prize, while if costs are convex, it might be optimal to distribute the budget amongst several prizes.[6] In M-S's theoretical contests, equilibrium effort functions are continuous functions of the abilities of the workers. In the lab, though, individual effort appears to follow a discontinuous step function in which low-ability workers drop out and exert zero or low effort, while

high-ability workers over-exert themselves. This trend leads to the bifurcation of efforts described above.

More relevant to the current inquiry, when Mueller and Schotter aggregate their data across individuals, efforts appear continuous. The observed bifurcation of efforts is obscured at the aggregate level. This is an example of aggregate results giving a distorted view of what is occurring at the individual level.

To see this more starkly, consider Figure 3.1 from Mueller and Schotter (2010).

Figure 3.1 presents data from the eight treatments of Mueller and Schotter (2010) aggregated over all subjects and all periods in each treatment. In the theory, a subject's effort in any given period of the experiment should be a function of the (random) ability of the subject in that period. The solid line in each figure represents the predictions of the theory, while each dot represents the mean effort chosen for any given ability level realized by subjects in that treatment.

Note that the equilibrium effort function of the theory is continuously decreasing in the ability of the subjects. Furthermore, note that, on average, the data suggest that behavior is predicted by the theory. Effort levels appear to be continuous in ability and basically attracted to the theoretical bid function, with a few exceptions. So as far as we can see, this theory is supported by the data if we aggregate in this manner.

This aggregation masks substantial deviation of the behavior of individual subjects from the predictions of the theory, however. Consider Figure 3.2, which represents a sample of individual bid functions taken from Mueller and Schotter.

In Figure 3.2 we see the dotted line representing the prediction of the theory while the dots again present the mean effort of the given subject when receiving a given ability level. Note that these individual bid functions tell a very different story from the aggregate functions presented in Figure 3.1. Here it is clear that subjects basically follow a step function in their bidding behavior. For high ability levels (represented by low numbers on the horizontal axis)—that is, ability levels below a critical threshold—subjects basically exert high effort levels. For ability levels above the threshold, they drop out. This qualitative feature of individual behavior is obscured by aggregation, where a composition effect makes the aggregate effort function appear continuous.

Some people might claim that this discrepancy (between aggregate and individual behavior) is of no practical importance, since what is important for organization design is whether the behavior of the organization (its aggregate behavior) is consistent with the theory used for its design. A risk-neutral manager will only care about aggregate effort and output and not how that output was achieved. This claim is misguided, however, for two reasons. First, the test performed by Mueller and Schotter did not investigate the comparative-static behavior of subjects over different contest parameters. While the theory would predict that behavior changes continuously over the space, it might be that certain configurations result in a disproportionate number of dropouts which would spoil the organizational atmosphere. Second, Mueller and Schotter used a design where subjects received

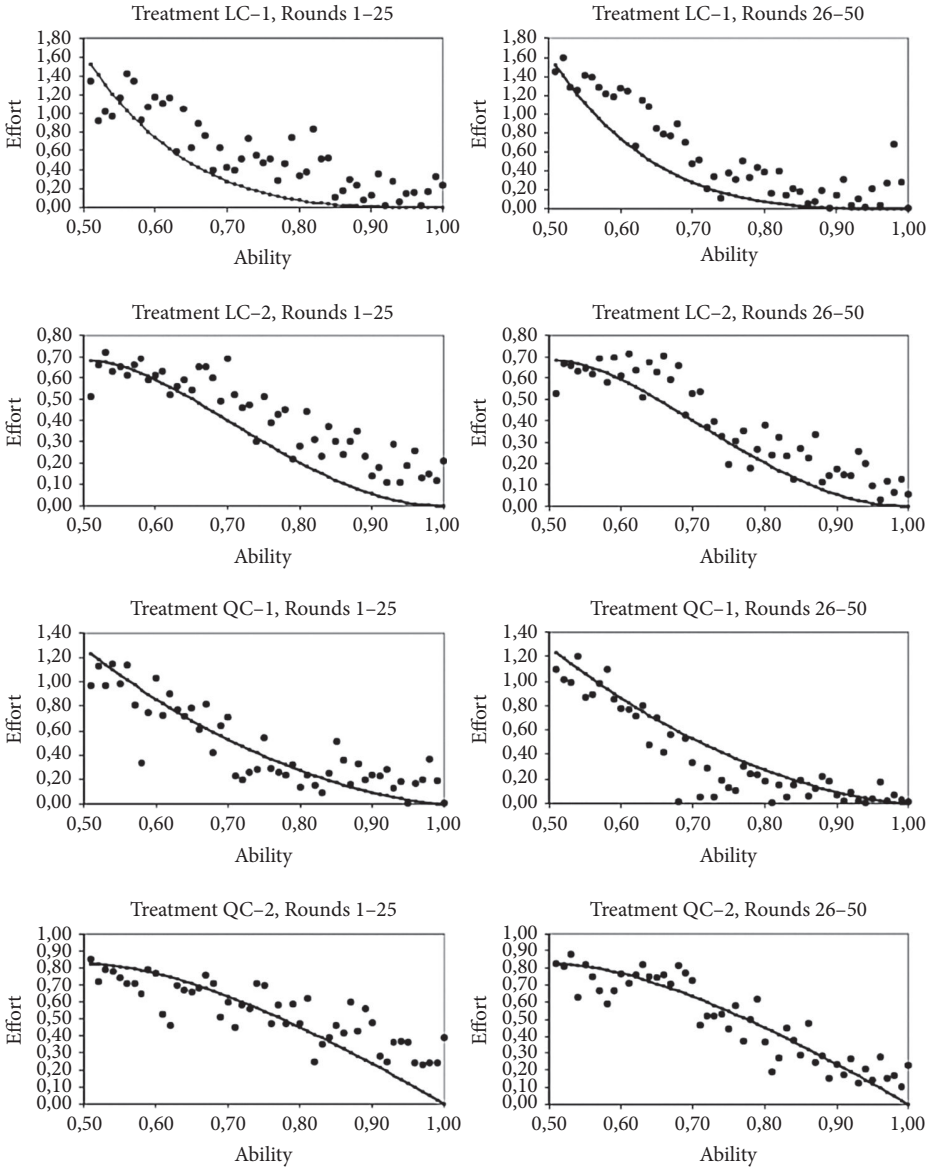


Figure 3.1. Average observed (dots) and optimal (solid line) effort functions in the first and the second half of the experiment.

a new random ability level each period, so that subjects who are low-ability contestants in one period may be high-ability in the next. If ability had been constant over the entire experiment—a design that more closely resembles the real world, if not the assumptions of the theory—then low-ability subjects that drop out might do so permanently. This permanent dropout might cause high-ability subjects to also lower their effort, leading to a potentially perpetual effort cycle.

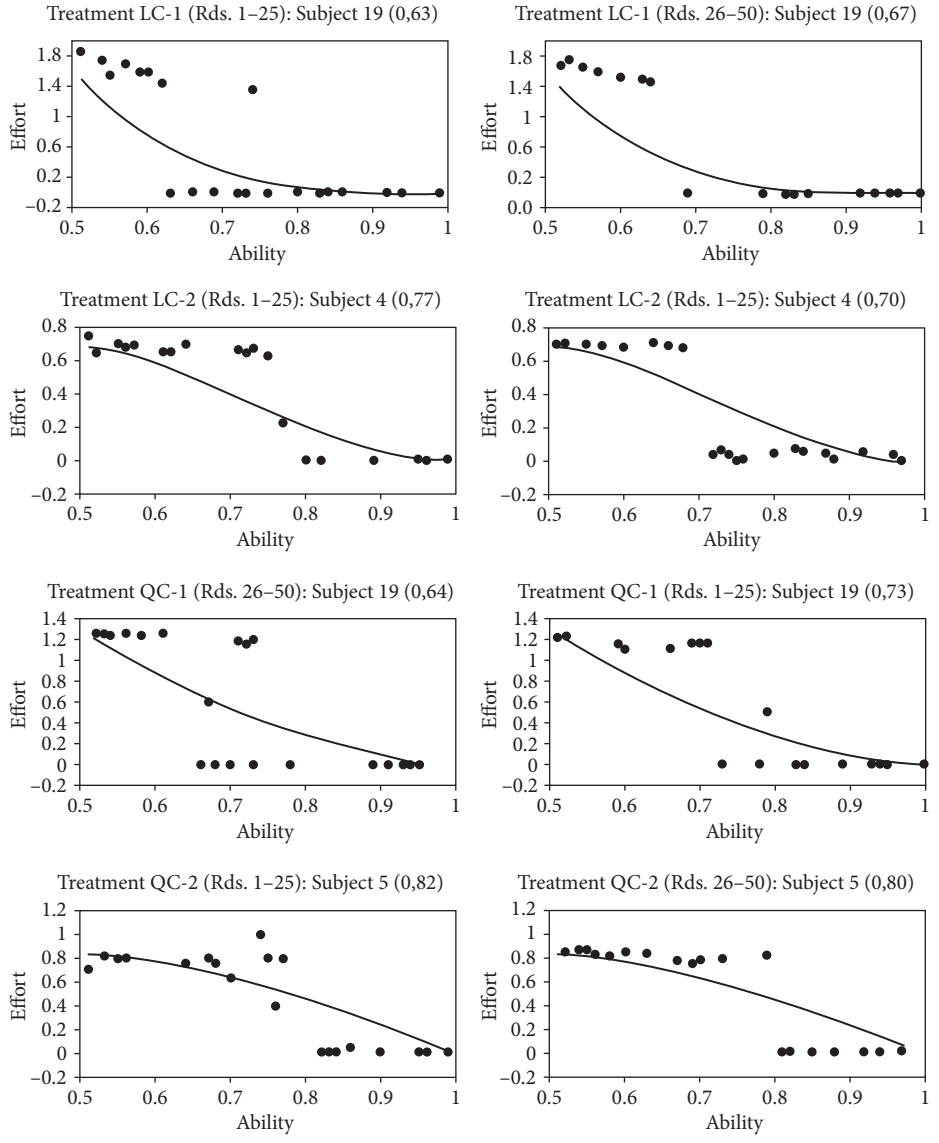


Figure 3.2. Examples of individual behavior (optimal solid line, observed dots).
***Note*: Cutoff levels in parentheses.**

A similar situation characterizes the learning literature, where various models are posited as describing learning behavior of subjects on the individual level yet are tested by aggregating data across subjects and time. For example, Erev and Roth (1998) define reinforcement models on the individual level yet aggregate over time in testing them, while Camerer and Ho's (1999) EWA model is also written on the individual level yet tested on the aggregate level. If such aggregation masks the same type of behavior as exists in the M-S experiment, then one would have to reevaluate the conclusions made there.

What Features of Data Present the Proper Test of the Theory?

This section will use learning models to illustrate how experimental data should and should not be used to test theories. Consider Figures 3.3a–3.3d, taken from unpublished data of Nyarko and Schotter (2000).

These figures present the period-by-period choices of individual subjects playing the game presented in Table 3.3. Subjects chose between two strategies, Green and Red.

In each figure we have placed the round of the experiment on the horizontal axis and the probability with which the subjects chose the Red strategy on the vertical. X's mark the pure strategy chosen in any given round, while circles indicate the prediions of the EWA model for that round. Dashed lines indicate the predictions of the Erev–Roth (Erev and Roth, 1998) reinforcement learning model, while the solid line indicates the prediction of the Nyarko and Schotter (2000) stated-belief learning model. The straight line indicates the static equilibrium prediction for the use of the Red strategy for the Row player.

Figures 3a–3d present the results for four subjects in the pure-strategy experiments where subjects could only use pure strategies.[7] What is striking about all of these diagrams is the failure of both the EWA and Reinforcement models to track the period-to-period movements of actions for individuals. Neither the EWA nor the Reinforcement models capture this movement. To the extent that they fit the

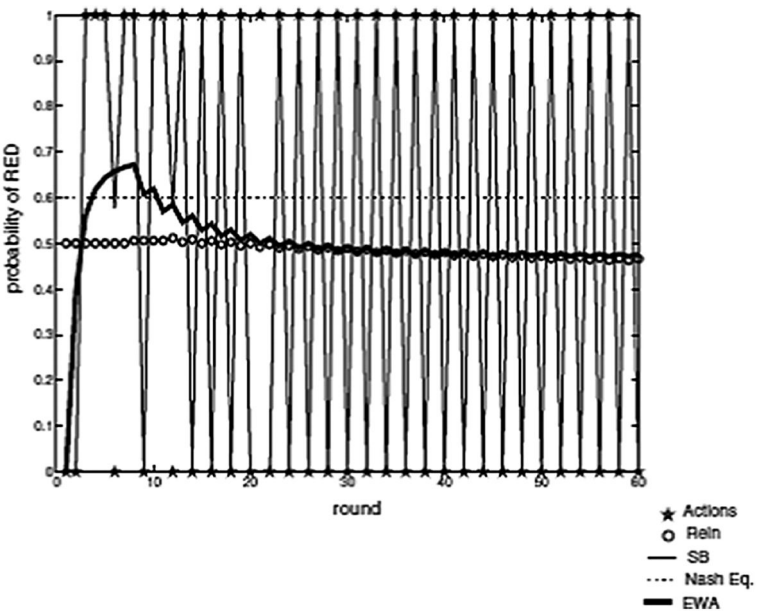


Figure 3.3a. Model predictions Experiment 1, Player 2.

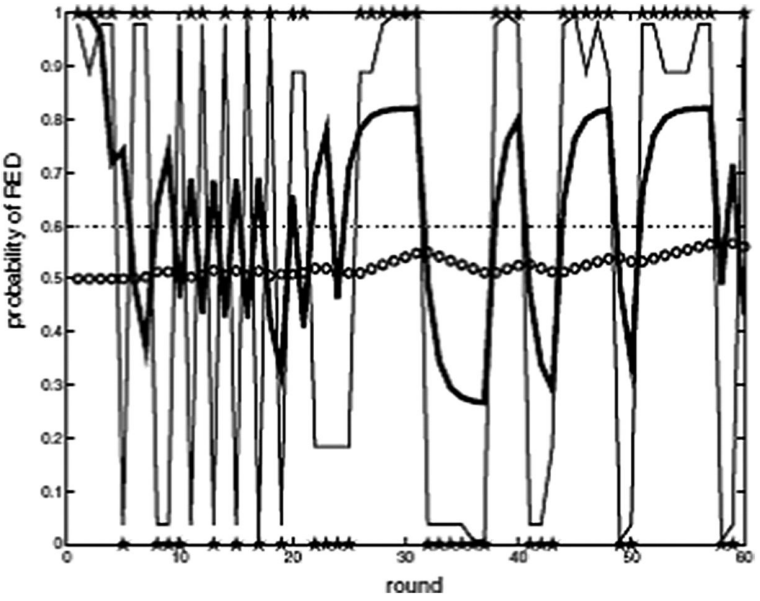


Figure 3.3b. Model predictions Experiment 1, Player 4.

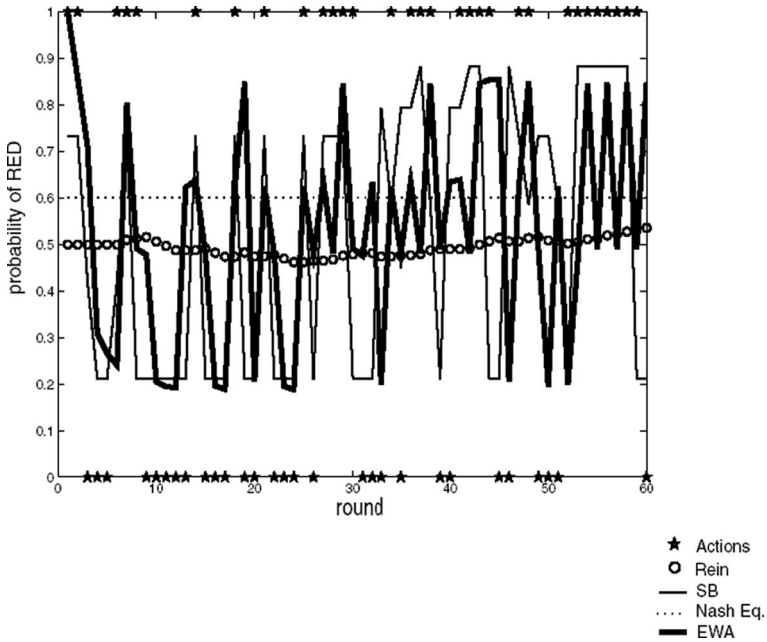


Figure 3.3c. Model predictions Experiment 1, Player 5.

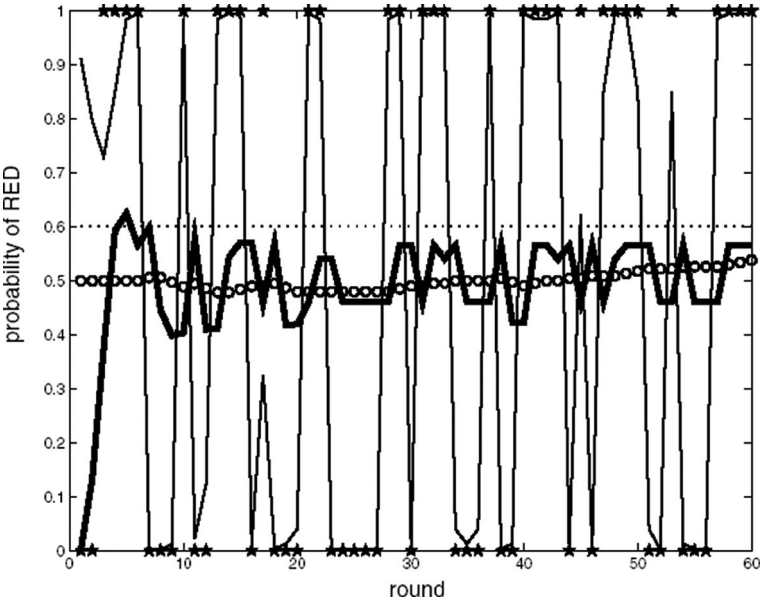


Figure 3.3d. Model predictions Experiment 1, Player 9.

Table 3.3. Nyarko-Schotter (2002) Game

	Green	Red
Green	6,2	3,5
Red	3,5	5,3

data, these models do so by passing a relatively straight line through a gyrating time series in an effort to minimize the errors in prediction being made. In contrast, the Stated-Belief model captures the qualitative movements of the data quite well. It predicts the abrupt changes seen in the data with a fair amount of precision because changes in actions in some sense mimic changes in beliefs, which are considerable.

From these figures we can see that each theory must be judged on two criteria. The first is that level calibration indicates how well the theory fits the data using a standard goodness-of-fit measure—that is, the mean squared deviation. The second is change calibration which indicates how well the theory captures the qualitative features of the data—in this case, the first difference of the time series. In most econometric exercises using laboratory data, authors restrict themselves to level calibration, judging models by their ability to match choice levels while ignoring their ability to predict changes from period to period. Ignoring change calibration leaves out a key qualitative feature of behavior, namely, behavioral volatility. Obviously, both features are relevant and need to be explained by any successful theory.

The data generated by these experiments offer observations on the time paths of chosen actions for subjects making choices over 60 rounds each. To compare the results of the three models mentioned above, Nyarko and Schotter (2000) constructed for each player and each theory two indices which describe how well the theory is calibrated to the levels and changes in the data. Our first measure is simply a mean MSD score calculated individual by individual in a particular experiment. In other words, the MSD score for individual i and model j, $j = 1, 2, 3$ is

$$\text{MSD}_i^j = \left(\frac{1}{N} \sum [p^j_{pred}(\text{Red}) - p_{act}(\text{Red})]^2 \right)^{1/2},$$

where N is the number of observations for individual i, $p^j_{pred}(\text{Red})$ is the predicted probability of Red being chosen by model j, and $p_{act}(\text{Red})$ is the actual probability that Red was chosen.

This equation determines your standard calibration score measuring the data's goodness of fit. As Figures 3.3a–3.3d indicate, however, such a measurement fails to capture the movement in the data. A model may be well-calibrated in terms of levels but achieve its good fit by passing a relatively flat time series of predictions through a constantly moving time series of actions. To correct for this, we take the first difference of both the actual choices made by our subjects and each theory's predictions. These first differences record the actual period-by-period change in the choices of subjects and predictions made about these changes. Comparing these two time series indicates whether a model predicts changes in the behavior of subjects and not just levels.

Formally, let a_{it} be the action chosen by subjects i in period t of the experiment he or she is engaged in. In the 2×2 experiments we are discussing, a_{it} will denote the probability weight placed on the Red strategy. $\Delta^a_{it} = a_{it} - a_{it-1}$ represents the change in the choice of subject i between period t and $t-1$. In the pure-strategy experiment studied here, Δ^a_{it} can take the values $\{-1, 0, +1\}$. Similarly, let $a^{pred(j)}_{it}$ denote the predicted probability weight placed on the Red strategy by learning model j, $j = 1, 2, 3$. $\Delta^{pred(j)}_{it} = a^{pred(j)}_{it} - a^{pred(j)}_{it-1}$ represents the change in the predictions of model j about the actions of subject i between period $t-1$ and t.

To compare learning models on the basis of whether they can predict the changes in behavior observed in an accurate manner, we propose simply to use the MSD metric on this first difference data as follows:

$$\text{MSD}_i^j = \left(\frac{1}{N} \sum [\Delta^{pred(j)}_{it} - \Delta^a_{it}]^2 \right)^{1/2}.$$

Hence, for any individual and any model, we have two goodness-of-fit measures, one measuring levels and one measuring changes in levels.

The results of these exercises are presented in Figures 3.4 and 3.5 using data from Experiment 1 of Nyarko and Schotter (2002). In that experiment, subjects

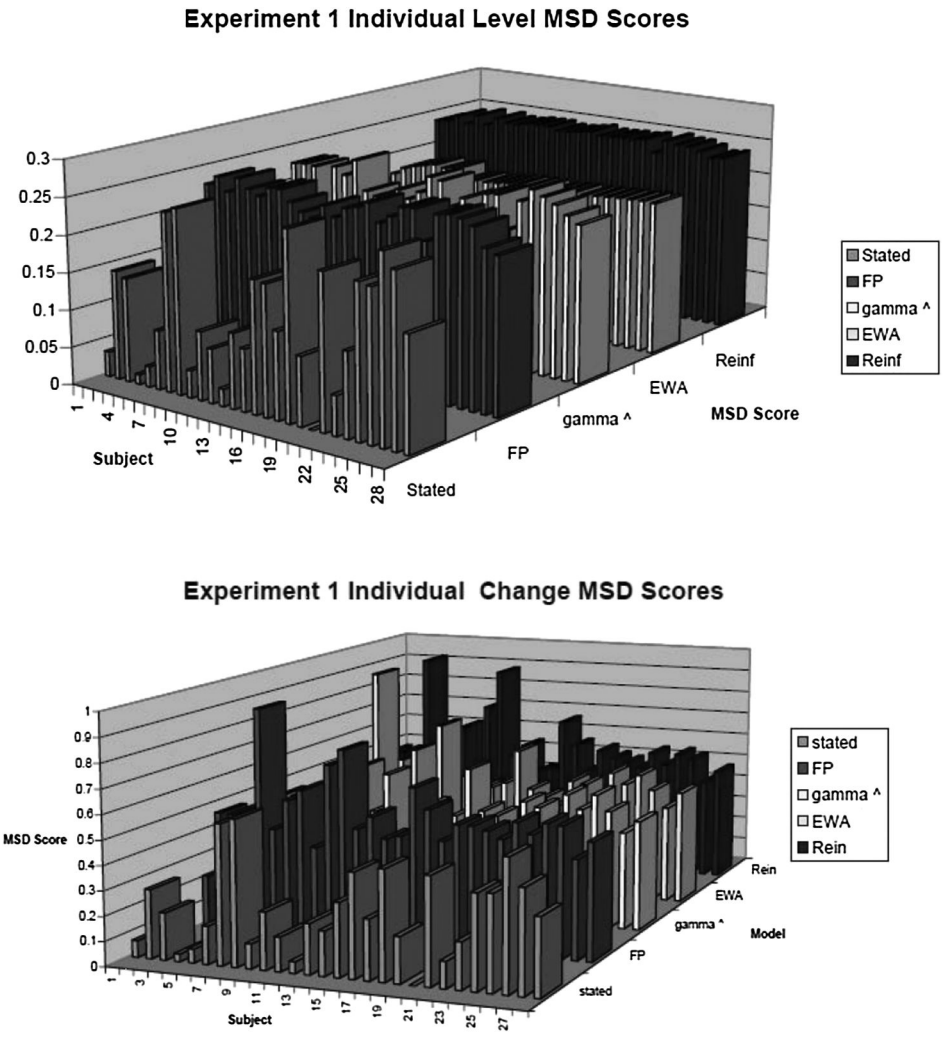


Figure 3.4. Individual level MSD scores from Nyarko and Schotter (2000).

were matched with the same partner for 60 periods. Their beliefs were elicited period by period, and their pure strategy choice was chosen. The prediction of four models are presented: EWA, reinforcement learning, and three belief-learning models, each using a different method of defining beliefs, including fictitious play (FP), stated beliefs (BS), and a third method whose definition need not concern us here.

As we can see, the stated-belief model of Nyarko and Schotter fits the data best in terms of level calibration. More importantly, it also fits the data best when we take into account the first difference in the data and therefore best characterizes the period-to-period changes in the behavior of subjects. This change calibration is important, in my view, because we should not consider a theory to have captured a

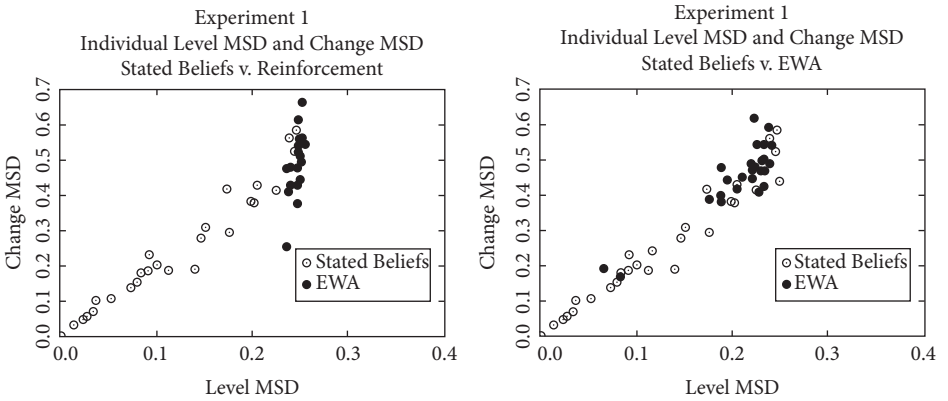


Figure 3.5. Level and change calibration MSD scores from Nyarko and Schotter (2000).

data set if it passes a relatively smooth and unchanging time series through a data set replete with volatile period-to-period gyrations. The stated belief model picks up on these changes because the elicited belief series underlying the model is itself very volatile. When plugged into a stochastic choice model, it predicts volatility in actions. The EWA and Reinforcement models are relatively sluggish and simply try to minimize mean squared deviations in the data by tracing the mean.

In summation, if we are to test theories with respect to their goodness of fit, the question of what we are fitting the data to becomes relevant. While most theories are fit to the levels of the choice variables in the data set, doing so may obscure other important features of the data, such as their volatility. Moreover, the validation conferred on a theory by a likelihood measure might be misleading if it is achieved by fitting the theory to the data in a manner that does not capture the qualitative features of the data.

Conclusion

This chapter has tried to answer some simple questions about the relationship between economic theory and experiments. It has argued that what makes economic theory useful as a guide to experiments is the structure that it provides for an examination of how people behave in economic situations. I have argued that one should proceed under the maintained hypothesis that theory is wrong, but hopefully wrong in interesting ways—that is, in ways that lead to a fuller description of behavior than was allowed for in the axioms underlying the theory—ways that will lead to new and better theories. I have also investigated the distinction between predictive theories and explanatory ones and pointed out that the original mission of economic theory, as spelled out by Friedman, was to generate predictive theories that are definable a priori—that is, before the data of the experiment is generated, as

opposed to explanatory theories that offer ex post explanations of data already generated. Finally, I have argued that given the constraints placed on social-scientific theory, it is probably best not to expect the lab to be able to do more than test the predictions of a theory and its comparative-static (qualitative) predictions.

One thing remains clear, however: Experimentation in economics is firmly established. It has provided the profession with a tool for testing theory without the need for heavy inference of the type done in field work. The beauty of experiments is that they allow an investigator to control what he/she needs to control and observe what he/she needs to observe, which allows for simplicity in the testing of models. The complicated identification strategies needed in field work can be replaced with the judicious use of experimental treatments that cleverly isolate variables of interest.

NOTES

1. See Levine and Zheng (Chapter 2, this volume) for a discussion of these assumptions. See also Schotter and Sopher (2007) for a discussion of how non-subgame-perfect equilibria with positive amounts offered by the Proposer may be selected as solutions to the Ultimatum Game.

2. The original Güth study showed that subjects were fully capable of performing backward induction in a more complex game, suggesting that they knew the subgame perfect equilibrium in the Ultimatum Game but rejected it based on other factors.

3. It is not enough to claim ex post that experimentalists have misinterpreted a theory, or else that if one allows for other preference assumptions or other solution concepts, the theory is resurrected. The theorist must determine the predictions of his theory before the experimental test, and not after the data have arrived. While we now can look at Ultimatum Game data and accuse the originators of naiveté, I think that they were testing the predictions of the prevailing theories at the time the experiment was run.

4. I'd like to thank Chris Flinn for this formalization.

5. See Fréchette, (Chapter 17, this volume) for the do's and don'ts of this approach.

6. Harbring and Irlenbusch (2003) report the effect of varying the prize structure in experimental rank order tournaments à la Lazear and Rosen (1981). They show, among other things, that average effort increases with a higher share of winner prizes.

7. Another set of experiments were run where subjects could use mixed strategies, but we will not comment on them here.

REFERENCES

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* **21**:503–546.

Baraji, P. and A. Hortaçsu. 2005. Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data. *Journal of Political Economy* **113**:703–741.

Bolton, G. E. and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* **90**(1):166–193.

Brown, M., C. Flinn, and A. Schotter. 2011. Real-Time Search in the Laboratory and the Market. *American Economic Review* **101**(2):948–974.
Camerer, C. F. and T. H. Ho. 1999. Experience-Weighted Attraction (EWA) Learning in Normal-Form Games. *Econometrica* **67**:827–874.
Camerer, C. F., T. H. Ho, and J. K. Chong. 2004. A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics* **119**(3):861–898.
Costa-Gomes, M. A., V. P. Crawford, and B. Broseta. 2001. Cognition and Behavior in Normal-Form Games: An Experimental Study. *Econometrica* **69**(5):1193–1235.
Crawford, V. P. and N. Iriberri. 2007. Fatal Attraction: Salience, Naivete, and Sophistication in Experimental "Hide-and-Seek" Games. *American Economic Review* **97**:1731–1750.
Ellsberg, D. 1961. Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics* **75**(4):643–669.
Erev, I. and A. Roth. 1998. Prediction How People Play Games: Reinforcement Learning in Games with Unique Strategy Equilibrium. *American Economic Review* **88**:848–881.
Feyerabend, P. 1993. *Against Method*. London: Humanities Press.
Fehr, E. and K. M. Schmidt. 1999. A Theory Of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* **114**(3):817–868.
Flinn, C. J. and J. J. Heckman. 1982. New Methods for Analyzing Structural Models of Labor Force Dynamics. *Journal of Econometrics* **18**:115–168.
Friedman, M. 1953. *Essays in Positive Economics*. Chicago: University of Chicago Press.
Goeree J., C. Holt, and T. Palfrey. 2002. Quantal Response Equilibrium and Overbidding in Private-Value Auctions. *Journal of Economic Theory* **104**(1):247–272.
Güth, W., R. Schmittberger, and B. Schwarze. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization* **3**(4):367–388.
Harbring, C. and B. Irlenbusch. 2003. An Experimental Study on Tournament Design. *Labour Economics* **10**(4):443–464.
Ho, T. H., X. Wang, and C. Camerer. 2008. Individual Differences in the EWA Learning with Partial Payoff Information. *Economic Journal* **118**:37–59.
Kahneman, D. and A. Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **XLVII**:263–291.
Lazear, E. P. and S. Rosen. 1981. Rank-Order Tournaments as Optimum Labor Contracts. *The Journal of Political Economy* **89**(5):841–864.
Moldovanu, B. and A. Sela. 2001. The Optimal Allocation of Prizes in Contests. *American Economic Review* **91**(3):542–558.
Mueller, W. and A. Schotter. 2010. Workaholics and Dropouts in Organizations. *Journal of the European Economic Association* **8**(4):717–743.
O'Neill, B. 1987. Nonmetric Test of the Minimax Theory of Two-Person Zerosum Games. *Proceedings of the National Academy of Sciences* **84**(7):2106–2019.
Nyarko, Y. and A. Schotter. 2000. Comparing Learning Models with Ideal Micro-Experimental Data Sets. Mimeo, New York University.
Nyarko, Y. and A. Schotter. 2002. An Experimental Study of Belief Learning Using Elicited Beliefs. *Econometrica* **70**:971–1005.
Rapoport, A. and R. B. Boebel. 1992. Mixed Strategies in Strictly Competitive Games: A Further Test of the Minimax Hypothesis. *Games and Economic Behavior* **4**(2):261–283.
Roberts, S. and H. Pashler. 2000. How Persuasive Is a Good Fit? A Comment on Theory Testing. *Psychological Review* **107**:358–367.
Rubinstein, A. and A. Tversky. 1993. Naïve Strategies in Zero-Sum Games. Sackler Institute of Economic Studies, Tel Aviv University Working Paper 17–93.

Rubinstein, A., A. Tversky, and D. Heller. 1996. Naïve Strategies in Competitive Games. In *Understanding Strategic Interaction—Essays in Honor of Reinhard Selten*, ed. W. Albers, W. Güth, P. Hammerstein, B. Moldovanu, and E. van Damme, 394–402. Berlin: Springer-Verlag.

Schotter, A. 2007. Strong and Wrong: The Use of Rational Choice Theory in Experimental Economics. *Journal of Theoretical Politics*, **18**(4):498–511.

Schotter, A. 2008. What's So Informative about Choice? In *The Foundations of Positive and Normative Economics: A Handbook*, eds. A. Caplin and A. Schotter. New York: Oxford University Press.

Schotter, A. and B. Sopher. 2007. *Advice and Behavior in Intergenerational Ultimatum Games: An Experimental Approach*. *Games and Economic Behavior* **58**(2):365–393.

Stahl, D. O. and P. W. Wilson. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior* **10**(1):218–254.

von Neumann, J. and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

CHAPTER 4

ENHANCED CHOICE EXPERIMENTS

ANDREW CAPLIN AND MARK DEAN

Introduction

Experiments that record more than standard choice data can improve our understanding of choice itself. In this chapter, we illustrate the advantage of such "enhanced" choice data with an example. We describe an experiment which, in addition to recording the final decision made by subjects, incentivizes choices that are made in the prior period of contemplation. We show how the resulting data provide insight into how people search for information on the alternatives available to them. This, in turn, improves our understanding of the decision-making process and its final outcome, which stands alone as the subject of interest in standard choice experiments.

This experiment forms part of an ongoing research project in which we aim to enrich the modeling of search behavior while retaining the theoretical discipline inherent in the revealed preference approach to choice behavior. We outline in some detail the poster child for our approach detailed in Caplin and Dean (2011) and Caplin, Dean, and Martin (2011), henceforth CD and CDM respectively. CD introduce "choice process" data, which measures not only the final option that a decision maker (DM) selects, but also how their choice changes with contemplation time before a final decision is made. CDM describe the results of an experiment designed to elicit choice process data in the laboratory.

We first describe how choice process data can be used to test models of information search and choice. CD characterize a model of sequential, or "alternative-based" search (ABS), in which a DM searches through alternatives sequentially, at

any time choosing the best of those they have searched according to a fixed utility function. Such behavior is standard within economic models of price and wage search. While the ABS model is silent as when people stop searching, CD provide a refinement that describes a DM who searches until an object is identified with utility above a fixed reservation level, as in the satisficing model of Simon (1955). We call this refinement "reservation-based search" (RBS). This form of search is optimal in simple models of sequential search with psychic search costs. Importantly, neither the ABS model nor the RBS model provides testable implications for standard choice data, meaning that choice process data are crucial in any such test.

Next, we describe a set of experiments reported in CDM designed to evaluate the ABS and RBS models. In order to simplify these tests, CDM specialize to the case of known preferences by making the objects of choice particularly simple: amounts of money. In order to prevent the choice problem from becoming trivial, these amounts are expressed in algebraic form (e.g., seven plus three minus two) and are therefore difficult to decode. While the experimenter always knows which of these numerical expressions is largest and therefore most preferred, the DM must work to assess each alternative. In this environment it is easy to identify choice "mistakes," or cases in which a subject has failed to choose the best alternative in a choice set.

In order to test the ABS and RBS models, our experiment elicits choice process data. We obtain such data using an experimental design in which subjects' choices are recorded at a random point in time unknown to them, incentivizing them to always report their currently preferred alternative. We therefore gather information on the sequence of choice switches in the pre-decision period, which we interpret as choice process data. This represents a choice-based experiment constructed precisely to enrich our understanding of search behavior and imperfect information.

There are three key findings:

1. There is evidence in favor of the ABS model. Specifically, the vast majority of switches are in the direction of improvement, suggesting that chosen objects have been accurately assessed.
2. There is strong evidence in favor of the satisficing model. Many decision makers engage in sequential search that stops once a satisfactory level of reservation utility is achieved. These reservation levels are environmentally determined—changing with the size of the choice set and the complexity of each alternative.
3. There are interesting individual differences in search behavior that intermediate the impact of search order on choice. For example, those who tend to search lists from top to bottom fail to choose the best option if it is far down the list, while those who search less complex options first miss the best option if it is complex, even if it is high on the list.

The section entitled "Choice Alone" shows that the choice process data are essential for making inferences concerning the decision-making process and its

impact on choice. When we re-run the same experiments gathering information only on final choices, little can be inferred about the forces that underlie mistakes, and the empirical regularities that are uncovered by exploring the choice process get obscured. It is not possible to test the ABS and RBS models, nor extract data on reservation levels.

The section entitled "Methodology" comments on the underlying motivation for our research and the broader methodology. We are far from the first to design experiments to study behavior in the pre-decision period. What defines our approach is the focus on choice-based enhancements that can be incentivized in an experimental laboratory. We believe that the axiomatic approach of standard choice theory provides the most robust foundation for understanding decisions. In this sense, the work described herein fits with a broader research agenda of introducing "nonstandard" data yet retaining the modeling discipline that axiomatic modeling provides. Caplin and Dean (2008) and Caplin et al. (2010) outline a distinct application of this approach that jointly characterizes standard choice data and neuroscientific data on the dopamine system.

The section entitled "Concluding Remarks" outlines the immediate next steps in the agenda. In the longer run, we see research on the relationship between search and choice as of ever-increasing importance in the age of Google and of policy "nudges." How we learn and choose when complex options are presented in various different manners is a question that will be increasingly under the microscope in the years to come, and on which research of the form outlined herein may shed light.

Choice Process Data: Theory

Choice process data are designed to provide insight into search-based causes of mistakes. Rather than recording only the final decision, these data track how the choices that people make evolve with contemplation time. As such, choice process data come in the form of sequences of observed choices from any given set of options rather than comprising a single set of chosen options (the theory allows for indifference and therefore simultaneous selection of several elements from a set). To formalize, let X be a nonempty finite set of elements representing possible alternatives, with $\mathcal{X}$ denoting nonempty subsets of X. Let $\mathcal{Z}$ be the set of all infinite sequences from $\mathcal{X}$ with generic element $Z = \{Z_t\}_{t=1}^{\infty}$ with $Z_t \in \mathcal{X}/\emptyset$ all $t \geq 1$. For $A \in \mathcal{X}$, define $Z \in \mathcal{Z}_A \subset \mathcal{Z}$ if and only if $Z_t \subset A$ all $t \geq 1$.

Definition 4.1.

*A **choice process** (X, C) comprises a finite set X and a function $C : \mathcal{X} \to \mathcal{Z}$ such that $C(A) \in \mathcal{Z}_A \; \forall \; A \in \mathcal{X}$.*

Given $A \in \mathcal{X}$, choice process data assign not just final choices (a subset of A) but a sequence of choices, representing the DM's choices after considering the problem for different lengths of time. We let C_A denote $C(A)$ and let $C_A(t) \in A$ denote the tth element in the sequence C_A, with $C_A(t)$ referring to the objects chosen after contemplating A for t periods. Choice process data represent a relatively small departure from standard choice data, in the sense that all observations represent choices, albeit constrained by time.

The first model that CD analyze captures the process of sequential search with recall, in which the DM evaluates over time an ever-expanding set of objects, choosing at all times the best objects thus far identified. Choice process data have an alternative-based search (ABS) representation if there exists a utility function and a nondecreasing search correspondence for each choice set such that what is chosen at any time is utility-maximizing in the corresponding searched set.

Definition 4.2.

Choice process (X, C) *has an* ***ABS*** *representation* (u, S) *if there exist a utility function* $u : X \to \mathbb{R}$ *and a search correspondence* $S : \mathcal{X} \to \mathcal{Z}^{ND}$*, with* $S_A \in \mathcal{Z}_A$ *all* $A \in \mathcal{X}$*, such that*

$$C_A(t) = \arg \max_{x \in S_A(t)} u(x),$$

where $\mathcal{Z}^{ND} \subset \mathcal{Z}$ *comprises nondecreasing sequences of sets in* $\mathcal{X}$*, such that* $Z_t \subset Z_{t+1}$ *all* $t \geq 1$.

Given that final choice of x over y is unrevealing with incomplete search, the ABS characterization relies on an enriched notion of revealed preference. To understand the required enrichment, we consider behavioral patterns that contradict ABS. In doing this, we use the notation $C(A) = B_1; B_2; \ldots ; B_n!$ with $B_i \in \mathcal{X} \cap A$ to indicate that the sets $B_1 \ldots B_n$ are chosen sequentially from A, with B_n being the final choice. The following choice process data all contradict ABS.

$$C^{\alpha}\left(\{x, y\}\right) = x; y; x!$$
$$C^{\beta}\left(\{x, y\}\right) = x; \{x, y\}; y!$$
$$C^{\gamma}\left(\{x, y\}\right) = y; x!; C^{\gamma}\left(\{x, y, z\}\right) = x; y!$$
$$C^{\delta}\left(\{x, y\}\right) = y; x!;\ C^{\delta}\left(\{y, z\}\right) = z; y!;\ C^{\delta}\left(\{x, z\}\right) = x; z!$$

C^{α} contains a preference reversal: The DM first switches to y from x, suggesting that x is preferred to y. However, the DM then switches back to y, indicating that y is preferred to x. C^{β} involves y first being revealed indifferent to x, as x and y are chosen at the same time. Yet later y is revealed to be strictly preferred to x as x is dropped from the choice set. In C^{γ} the direction in which preference is revealed as between y and x changes between the two element and three element

choice set. C^{δ} involves an indirect cycle, with separate two-element sets revealing x as preferred to y, y as preferred to z, and z as preferred to x.

As these examples suggest, the appropriate notion of strict revealed preference in the case of ABS is based on the notion of alternatives being replaced in the choice sequence over time. A DM who switches from choosing y to choosing x at some later time is interpreted by the ABS model as preferring x to y. Similarly, if we ever see x and y being chosen at the same time, it must be that the DM is indifferent between the two alternatives. Hence we capture the revealed preference information implied by the ABS model in the following binary relations.

Definition 4.3.

Given choice process (X, C), *the symmetric binary relation* $\sim$ *on* X *is defined by* $x \sim y$ *if there exists* $A \in \mathcal{X}$ *such that* $\{x, y\} \subset C_A(t)$ *some* $t \geq 1$. *The binary relation* $\succ^C$ *on* X *is defined by* $x \succ^C y$ *if there exists* $A \in \mathcal{X}$ *and* $s, t \geq 1$ *such that* $y \in C_A(s)$, $x \in C_A(s+t)$ *but* $y \notin C_A(s+t)$.

For a choice process to have an ABS representation, it is necessary and sufficient for the revealed preference information captured in $\succ^C$ and $\sim$ to be consistent with an underlying utility ordering. The CD characterization of ABS therefore makes use of Lemma 4.1, a standard result which captures the conditions under which an incomplete binary relation can be thought of as reflecting some underlying complete pre-order. Essentially, we require the revealed preference information to be acyclic.

Lemma 4.1.

Let P *and* I *be binary relations on a finite set* X, *with* I *symmetric, and define* PI *on* X *as* $P \cup I$. *There exists a function* $v : X \to \mathbb{R}$ *that* ***respects*** P *and* I:

$$xPy \Longrightarrow v(x) > v(y),$$
$$xIy \Longrightarrow v(x) = v(y),$$

if and only if P *and* I *satisfy* ***OWC*** *(only weak cycles): Given* $x_1, x_2, x_3, \ldots, x_n \in X$ *with* $x = x_1 PI x_2 PI x_3 \ldots PI x_n = x$, *there is no* k *with* $x_k P x_{k+1}$.

Armed with this result, CD establish that the key to existence of an ABS representation is for $\succ^C$ and $\sim$ to satisfy OWC.

Theorem 4.1.

Choice process (X, C) *has an ABS representation iff* $\succ^C$ *and* $\sim$ *satisfy OWC.*

This condition is closely related to the standard strong axiom of revealed preference. It is this condition that reduces to the improvement condition that is tested

in the experiment described in the section entitled "The Experiment." The set of equivalent representations of a choice process for which $\succ^C$ and $\sim$ satisfy OWC involve (a) the utility function v respecting $\succ^C$ and $\sim$ on X and (b) the search correspondence S including at least all objects which have been chosen from all sets A at times $s \leq t$, where permissible additional elements that have utility are strictly below that associated with chosen objects according to v. Hence the more switches there are between objects in the choice process, the more restricted is the set of utility functions that can form part of an ABS representation.

Since the ABS model says nothing about the stopping rule for search, CD augment it with a simple "reservation utility" stopping rule in which search continues until an object is found which has utility above some fixed reservation level, whereupon it immediately ceases. The key to the empirical content of this stopping rule is that one can make inferences as to objects that must have been searched even if they are never chosen. Specifically, in any set in which the final choice has below reservation utility, it must be the case that all objects in the set are searched. Hence final choices may contain revealed preference information. The RBS model embodies the concept of satisficing that Simon (1955) introduced in his pioneering model of bounded rationality, in which he suggested that decision makers do not optimize but rather search until they achieve a "satisfactory" (or reservation) level of utility.

Intuitively, an RBS representation is an ABS representation (u, S) in which a reservation level of utility ρ exists, and in which the above-reservation set $X_u^\rho = \{x \in X | u(x) \geq \rho\}$ plays an important role in the search process. Specifically, search stops if and only if an above-reservation item is discovered. In order to capture this notion formally, CD define $C_A^L = \lim_{t\to\infty} C_A(t)$, as the final choice the DM makes from a set $A \in \mathcal{X}$ as well as limit search sets $S_A^L \equiv \lim_{t\to\infty} S_A(t) \in \mathcal{X}$. Note that, for finite X, the existence of an ABS representation guarantees that such limits are well defined.

Definition 4.4.

Choice process (X, C) *has a* ***reservation-based search*** *(**RBS**) representation* (u, S, ρ) *if* (u, S) *form an ABS representation and* $\rho \in \mathbb{R}$ *is such that, given* $A \in \mathcal{X}$,

R1 *If* $A \cap X_u^\rho = \emptyset$, *then* $S_A^L = A$.
R2 *If* $A \cap X_u^\rho \neq \emptyset$, *then:*
(a) *there exists* $t \geq 1$ *such that* $C_A(t) \cap X_u^\rho \neq \emptyset$;
(b) $C_A(t) \cap X_u^\rho \neq \emptyset \Longrightarrow S_A(t) = S_A(t+s)$ *all* $s \geq 0$.

Condition R1 demands that any set containing no objects above reservation utility is fully searched. Condition R2(a) demands that search must at some point

uncover an element of the above-reservation set if present in the feasible set. Condition R2(b) states that search stops as soon as reservation utility is achieved.

As with the ABS model, the key to characterizing the RBS model is to understand the corresponding notion of revealed preference. As RBS is a refinement of ABS, it must be the case that behavior that implies a revealed preference under ABS also does so under RBS. However, the RBS model implies that some revealed preference information may also come from final choice, with sets that contain only below-reservation utility objects being completely searched.

The following cases that satisfy ABS but not RBS illustrate behaviors that must be ruled out:

$$C^{\alpha}\left(\{x,y\}\right) = x; y!;\ C^{\alpha}\left(\{x,z\}\right) = x!;\ C^{\alpha}\left(\{y,z\}\right) = z!$$
$$C^{\beta}\left(\{x,y\}\right) = x; y!;\ C^{\beta}\left(\{x,y,z\}\right) = x!$$

In the first case, the fact that x was replaced by y in $\{x,y\}$ reveals the latter to be preferred and the former to be below reservation utility. (otherwise search must stop as soon as x is found). Hence the fact that x was chosen from $\{x,z\}$ reveals z to have been searched and rejected as worse than x, making its choice from $\{y,z\}$ contradictory. In the second, the fact that x is followed by y in the choice process from $\{x,y\}$ reveals y to be preferred to x and reveals x to have utility below the reservation level. The limit choice of x from $\{x,y,z\}$ therefore indicates that there must be no objects of above-reservation utility in the set. However, this in turn implies that the set must be fully searched in the limit, which is contradicted by the fact that we know that y is preferred to x and yet x is chosen.

In terms of ensuring existence of an RBS representation, the critical question is how to identify all objects that are revealed as having below-reservation levels of utility. As in the above cases, we know that an object must have utility below the reservation level if we see a DM continue to search even after they have found that object. CD call such an object nonterminal. Furthermore, we know that an object must be below reservation utility if, in some choice set, a directly nonterminal element is finally chosen instead of that object. CD define the union of this class of object and the nonterminal objects as indirectly nonterminal.

Definition 4.5.

Given choice process (X, C)*, define the nonterminal set* $X^N \subset X$ *and the indirectly nonterminal set* $X^{IN} \subset X$ *as follows:*

$$X^N = \{x \in X | \exists A \in \mathcal{X} \text{ s.t. } x \in C_A(t) \text{ and } C_A(t) \neq C_A(t+s) \text{ some } s, t \geq 1\},$$
$$X^{IN} = X^N \cup \{x \in X | \exists A \in \mathcal{X},\ y \in X^N \text{ with } x, y \in A \text{ and } y \in C_A^L\}.$$

Under an RBS representation, final choices in sets with below reservation utility objects contain revealed preference information: When choice is made from

two objects $x, y \in X$ either of which is indirectly nonterminal, then we can conclude that the chosen object is preferred. To see this, suppose that y is indirectly nonterminal, hence has below reservation utility. In this case if it is chosen over x, it must be that x was searched and rejected. Conversely, suppose that x is chosen over y. In this case either x is above reservation, in which case it is strictly preferred to y, or it is below reservation, in which case we know that the entire set has been searched, again revealing x superior. This motivates the introduction of the binary relation $\succ^L$ on X which gets united with the information from $\succ^C$ to produce the new binary relation $\succ^R$ relevant to the RBS case.

Definition 4.6.

Given choice process (X, C)*, the binary relation* $\succ^L$ *on* X *is defined by* $x \succ^L y$ *if* $\{x \cup y\} \cap X^{IN} \neq \emptyset$*, and there exists* $A \in \mathcal{X}$ *with* $x, y \in A$*,* $x \in C_A^L$*, yet* $y \notin C_A^L$*. The binary relation* $\succ^R$ *is defined as* $\succ^L \cup \succ^C$*.*

The behavioral condition that is equivalent to the RBS model is that the revealed preference information obtained from $\succ^R$ and $\sim$ is consistent with an underlying utility function.

Theorem 4.2.

Choice process (X, C) *has an RBS representation iff* $\succ^R$ *and* $\sim$ *satisfy OWC.*

The ABS and RBS models both treat search order as unobservable, and characterize the extent to which it is recoverable from choice process data. This makes it natural to develop stochastic variants, since there is no reason to believe that search from a given set will always take place in the same order. CD therefore generalize the deterministic model to allow for stochasticity. They do this by allowing each choice set to map onto a probability distribution over sequences of chosen objects. The resulting stochastic models turn out to be direct generalizations of their deterministic counterparts. CD show that the stochastic RBS model can capture anomalous choice behavior, such as status quo bias, stochastic choice, and general framing effects.

The Experiment

Having developed a theory of information search that could potentially explain choice "mistakes," our next task was to develop an experimental methodology that would allow us to test these models. These experiments are described in CDM.

The main simplification in the choice process experiment is that CDM make the utility function observable and identical across subjects. To accomplish this, the

objects of choice are amounts of money received with certainty. In order to make the choice problem nontrivial, each object is displayed as an arithmetic expression, a sequence of addition, and subtraction operations, with the value of the object equal to the value of the sum in dollars. The value of each alternative was drawn from an exponential distribution with $\lambda = 0.25$, truncated at \$35 (a graph of the distribution was shown in the experimental instructions).[1] Once the value of each object was determined, the operations used to construct the object were drawn at random.

Each round began with the topmost option on the screen selected, which had a value of \$0 and thus was worse than any other option. To elicit choice process data, subjects were allowed to select any alternative in the choice set at any time, changing their selected alternative whenever they wished. The alternative that the subject currently selected would then be displayed at the top of the screen. A subject who finished in less than 120 seconds could press a submit button, which completed the round as if they had kept the same selection for the remaining time. Typically, a subject took part in a single session consisting of two practice rounds and 40 regular rounds, and two recorded choices were actualized for payment, which was added to a \$10 show-up fee.

The key to the experimental design is the way in which subjects were incentivized. Rather than simply receiving their final choice, actualized choice was recorded at a random point in time unknown to the experimental subject. Specifically, subjects were instructed that at the end of the round, a random time would be picked from distribution between 1 and 120 seconds according to a truncated beta distribution with parameters $\alpha = 2$ and $\beta = 5$, and the selected alternative at this time would be recorded as the choice for that round.[2] At any given time, it is therefore optimal for the subject to have selected the alternative that they currently think is the best, as there is a chance that their current selection would be recorded as their choice. We therefore interpret the sequence of selections as choice process data.[3]

There were six treatments, differing in the complexity of choice object (3 or 7 addition and subtraction operations for each object) and the total number of objects (10, 20, or 40 alternatives) in the choice set. Figure 4.1 (from CDM) shows a 10-option-choice set with objects of complexity 3.

The experimental design creates an environment in which subjects' final choices were suboptimal. Averaging across all treatments, subjects fail to finally choose the best option 44% of the time. Failure rates vary from 11.4% for the size 10, low complexity (3 operations) treatment to 80.9% for size 40, high complexity (7 operations) treatment. These failures of optimality were also significant in terms of dollar amounts, with an average gap of more than \$8.00 between the finally chosen and the best option in the largest and most complex choice sets (size 40, complexity 7).

The potential for choice process data to shed light on the above losses derives from the fact that several switches are commonly observed in the pre-decision period. Most individuals do indeed change their selection with consideration time.

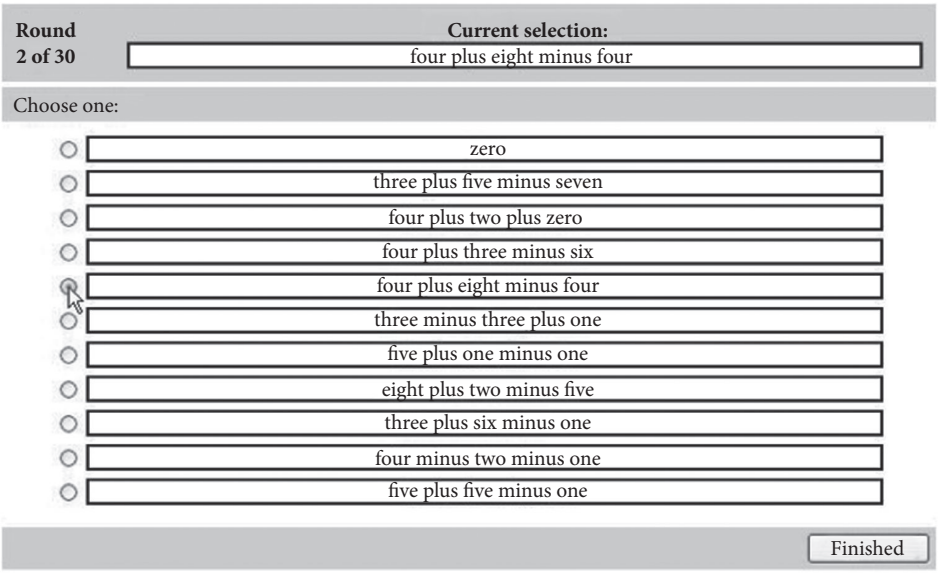


Figure 4.1. An example of a screen presented to experimental subjects. Subjects have to choose from the 10 alternatives on the screen. The value of each alternative is the value of the relevant sum. Subjects can select an alternative at any time by clicking on it, and they can change their choices as many times as they like.

This is a necessary condition for choice process data to contain more information than standard choice data alone.

Sequential Search

The first question that CDM consider is the extent to which switches in the pre-decision period are from lower to higher value alternatives (this corresponds to "alternative-based" search (ABS) in the general characterization above). Using a standard measure of the failure of consistency with revealed preference (Houtman and Maks, 1985) with measures of statistical power based on the alternative of random selection (as in Bronars (1987)), CDM show that, for the population as a whole, ABS does a good job of describing search behavior. They also identify "ABS types" by comparing each subject's HM index with the median HM index of the 1000 simulations of random data for that subject, which have exactly the same number of observations in each round. CDM classify as "ABS types" those subjects who have an HM index higher than the 75th percentile of their associated random data. Seventy-two out of 76 subjects fall into this category.

The prevalence of ABS types suggests that simple search theoretic explanations can help make sense of apparent mistakes. In large choice sets, people still recognize preferred objects and choose them when they come across them. However, their final choices may not be maximal because they do not search through all available alternatives.

Satisficing

CDM use choice process data to shed new light on satisficing behavior. They show that the RBS model describes the experimental data well at both the aggregate and individual level. At the aggregate level, for each treatment there exists possible reservation values such that, on average, subjects continue to search when currently selecting an alternative which is below this reservation level, but stop searching when holding a value above it. This is true even if the data are broken down by total number of searches. This means that a reservation level can be estimated for each treatment (as we describe below). The resulting estimated reservation values also do a good job of describing individual level data: Across all treatments, subjects behave in line with these reservation levels (i.e., stop searching when holding a value above this level but continue searching when holding a value below it) in approximately 77% of observations.

CDM use standard methods to estimate the reservation utility for each treatment. Specifically, all individuals in a given choice environment are assumed to have the same constant reservation value $\overline{v}$ and experience variability ε in this value each time they decide whether or not to continue search. Furthermore, this stochasticity is assumed to enter additively and to be drawn independently and identically from the standard normal distribution. To estimate the reservation level using choice process data, CDM consider each selection made by a subject as a decision node. Search is defined as continuing if a subject switches to another alternative after the current selection. Conversely, search is stopped at a decision node only if (a) the subject made no further selections and pressed the submit button and (b) the object they had selected was not the highest value object in the choice set. Reservation levels are estimated by maximizing the likelihood of observed decisions.

Not only do the estimated reservation levels do a good job of explaining individual behavior, they also shed light on why the number of mistakes vary from treatment to treatment. CDM find that reservation levels vary systematically with treatment. Reservation levels *decrease* as the objects of choice get more complicated. This explains why mistakes (measured as the gap in value between the chosen option and the best available option) are larger in more complicated choice sets. Reservation levels *increase* as the size of the choice set increases. This increase implies that subjects do better in absolute terms in larger choice sets, but the increase is not sufficient to prevent them from making larger mistakes in bigger choice sets.

While the theoretical interest of the satisficing model is clear, it is perhaps surprising that the experimental results of CDM offer such strong support to this stark model. A partial explanation may lie in the connection between the experimentally identified reservation stopping rules and optimal stopping rules in a model with a fixed psychic cost of search and independently drawn object valuations. CDM show that a fixed reservation strategy is optimal in this model. However there are some conflicts between this model and the experimental findings. Specifically, optimal

reservation levels are independent of the size of the choice set in the optimizing model, yet increasing in the experiment. Understanding this finding is a priority in ongoing research.

Search Order

Choice process data provide insight into the order in which people search through available objects, and this information can help predict when subjects will do badly in particular choice sets. In further experiments we analyzed two factors that can determine search order: screen position and object complexity. In order to explore both factors, we ran an additional experimental treatment which contained objects of varying complexity. This treatment contained choice sets of size 20, and the objects in each set varied in complexity from between one and nine operations. We ran the new treatment on 21 subjects for a total of 206 observed choice sets.

In the context of these experiments, we show that average search behavior has systematic patterns. On average, subjects search the screen from top to bottom: Screen position is higher for later searched objects. We show also that subjects tend to search from simple to complex objects. Perhaps more surprising is evidence of individual differences in the search patterns of individual subjects. Some subjects behave in a manner consistent with "Top–Bottom" (TB) search, while others are consistent with "Simple–Complex" (SC) search. The former are subjects whose search order takes them from the top to the bottom of the screen, while the latter are subjects whose search takes them from simple to complex objects.

The experiment reveals that the differences in search order impact final choices. In a round in which the highest-valued item is very short and occurs at the end of the list, TB searchers find it less often than do SC searchers. Conversely, when the highest-valued item is very long and occurs very early in the list, TB searchers find it more often than do SC searchers.

Choice Alone

Our central claim in this chapter is that our enhanced choice experiment helps us to understand choice behavior in a way that would not be possible using choice alone. In order to support this claim, we need to show two things. First, that our method for eliciting the enhanced choice data does not distort the behavior of our subjects to such an extent that we learn very little about standard choice environments. Second, that the additional data that we collected does in fact add to our understanding of choice.

To investigate these two points, CDM ran a "pure choice" version of the experimental design removing the incentives relating to the pre-decision period. The standard choice experiment made use of exactly the same treatments as the choice

process experiments: Choice sets contained 10, 20, or 40 alternatives, with the complexity of each alternative being either 3 or 7 operations. Moreover, exactly the same choice sets were used in the choice process and standard choice experiments. The subjects in the pure choice experiment took part in a single experimental session consisting of two practice rounds and between 27 and 36 regular rounds, drawn from all six treatments. At the end of the session, two regular rounds were drawn at random, and the subject received the value of the selected object in each round, in addition to a $10 show-up fee. Each session took about an hour, for which subjects earned an average $32. In total we observed 22 subjects making 657 choices.[4]

We find similar patterns of final choices in the pure choice and choice process environments. There are somewhat fewer mistakes in the pure choice experiment: Averaging across all treatments, subjects fail to select the best option 38% of the time, compared to 44% of the time in the choice process experiment. However, the comparative statics are similar in the two cases: Mistakes increase both with the size of the choice set and the complexity of alternatives, with failure rates varying from 7% for the size 10, low complexity (3 operations) treatment to 65% for the size 40, high complexity (7 operations) treatment. These failures of rationality remain significant in terms of dollar amounts, with average loss of $7.12 in the size 40, high complexity treatment. The pattern of mistakes was also similar between the pure choice and choice process settings: When CDM compare the distribution of final choices using Fisher's exact test, only 12 of the 60 choice sets have distributions that are significantly different at the 5% level.

To the extent that there is a difference in the quality of final choices between the choice process and pure choice treatments, it goes in the expected direction. The incentive to continue searching is higher in the standard choice experiment, since it is certain that any identified improvements will be implemented. The corresponding probability is less than one in the choice process experiment, and it falls toward zero as the 2 minutes come to an end. In this light, it is noteworthy how limited was the impact of the incentive changes induced by the choice process interface.

There is one other source of evidence on the similarity in decision making with and without the enhanced incentives in the choice process experiment. The experimental design allowed subjects in the pure choice experiment to select options prior to their final choice just as they could in the choice process experiment. The only difference was that in the standard choice experiment, there was no incentive for them to do so. CDM found that subjects still did record switches of their own volition even without incentives, that the resulting selections broadly satisfied ABS and RBS, and that reservation utilities exhibited the same qualitative patterns as in the incentivized experiment. Essentially all of the results above concerning the nature of the search and decision process from the choice process experiments are closely mirrored using data from the pre-decision period in the pure choice experiment despite the absence of incentives.

How much could we have learned about information search and choice had we observed only pure choice and ignored the pre-decision period? The answer is

very little, and much of it wrong. On the positive side, one would learn that choices are made more poorly in the larger and more complex decision sets. On the negative side, one would have no way of testing various explanations for what is behind these poor decisions. With choice data alone, one could not test the ABS or RBS model, nor make reliable inferences about reservation utility levels. If all one observed were final choices, then any data set can be explained perfectly by a model in which the reservation level is zero, and whatever is chosen is the first object searched. Thus it is infeasible to estimate reservation levels and compare how they change with the environment. The extra information in the choice process allows us to understand what it is that drives suboptimal choice and estimate otherwise hidden reservation parameters.

A second advantage of choice process data is that they allow us to recover revealed preference information in the case of incomplete search. In environments such as these, where all alternatives are not evaluated, the eventual choice of one object over another does not necessarily convey preference, as the decision maker may be unaware of the unchosen alternative. Standard choice data do not therefore reveal any preference. In contrast, when one has access to choice process data, the switch from one alternative to another *does* convey revealed preference information under the assumption of the ABS model, because the fact that the former alternative was chosen at one time indicates that the decision maker was aware of its existence.

Methodology

The research outlined above is part of a methodologically oriented agenda in the area of "neuroeconomics." This field has recently been the subject of much controversy concerning its definition and its substance. An initial salvo was fired by Camerer et al. (2005), who argued that neurological data would revolutionize our understanding of choice. Gul and Pesendorfer (2008) fired back hard with the claim that nonstandard data is essentially irrelevant to economics, which is interested only in the act of choice.

Following this harsh exchange, the center of the active debate on neuroeconomics concerns what are appropriate forms of nonstandard data to explore in order to better understand choice, as well as the extent to which these data need themselves to be modeled. Here there are many flowers that continue to bloom. Camerer (2008) outlines a very wide array of nonstandard data that are potentially interesting to those seeking to understand choice. Search in particular has been a major spur to the development of psychological data. Herbert Simon developed "protocol analysis" to augment choice data with highly structured vocalized descriptions of the decision-making process (Ericsson and Simon, 1984); time to decide has been the focus of much research (e.g., Armel et al. (2006) and Wilcox (1993)), as have the order of information search as revealed by

Mouselab (e.g., Payne et al. (1993), Ho et al. (1998), and Gabaix et al. (2006)), eye movements (e.g., Wang et al. (2010) and Reutskaja et al. (2011)), and neuroscientific observations.

The program of neuroeconomic research in which we are engaged, and to which the work outlined herein contributes, involves a particularly tight relationship between nonstandard data and economic theory. We see the tension between tightly constrained decision theory and massive volumes of new psychological data as potentially damaging to the social nature of the research enterprise (see Caplin (2008) for an in-depth exposition). The concern is that such open-ended constructs as decision making frames, mental accounts, and rules of thumb are flexible enough to account for any pattern of observations and are subjective enough to defy common definition. We believe that the key to avoiding this potential communication breakdown is to internalize the profound strength of the data theoretic ("axiomatic") approach introduced into economics by Samuelson (1938).

It is ironic that the axiomatic approach is traditionally seen as connected to a standard concept of choice among available alternatives. There is no necessary connection of this nature. Indeed it is our view that axiomatic methods are made ever more essential by data proliferation. Application of the axiomatic method ensures that new data earn their keep by opening up new observable phenomenon. In principle, axiomatic methods represent an ideal way of unifying psychologically sophisticated decision theory with experimental economics. It is also methodologically incoherent to argue that axiomatic methods apply best to a particular designated data set, comprising "standard" choices. We see no valid distinction between choice and nonchoice data. To take an extreme case, the pulse can be modeled as chosen just as much as can the standard choice of apples over oranges. While it may be that the former is more tightly constrained by physical laws, even this is debatable. After all, the goal of choice theory is to treat choice itself as mechanically as possible.

The first work in which we jointly characterize properties of standard choice data and nonstandard data relates to the neurotransmitter dopamine. In that context, Caplin and Dean (2008) identified the precise characteristics of the standard theory in which dopamine provides a reward prediction error signal, while Caplin et al. (2010) provided the corresponding experimental tests, which were broadly positive. The current work is the second example, but is in many ways more fundamental to the methodology. It takes full advantage of the researcher's freedom to specify nonstandard data that is experimentally observable, and for which a ready-made theory exists that is very close to standard choice theory. An extremely well-developed theory suggests that search is sequential and investigates optimal search, which is often of the reservation variety. Moreover, the very first breakdown of revealed preference relates to incomplete search, a point that was noted early on by Block and Marschak (1960) in their pioneering model of stochastic choice.

Our investigation of choice process data reflects the joining of natural streams: study of nonstandard data and axiomatic methods of choice theory. Interestingly, Campbell (1978) had previously developed a theory of this data tape with respect to

an early model of the decision-making procedure. In fact it is in many respects a natural data tape for a theorist, and ABS and RBS are natural first formulations of boundedly rational decision models.

Concluding Remarks

There are several obvious next steps in the research agenda related to the choice process. One such step is to join choice process data with additional observations on the search process, including mouse movements, time between switches, eye movements, and neurological measurements. We expect eye movements to be of particular value in helping us understand the nature of search. While less well-studied than standard choice behavior, opening up to these enriched observations may be very important in analyzing possible alternative modes of search, such as "characteristic-based" procedures in which objects are compared on a facet-by-facet basis.

With regard to applications, we are particularly interested in variants of the choice process model and experiment that give insight into financial decision making over the Internet. It is intuitively clear that most of us are incapable of making fully informed financial decisions and that the mode of presentation can substantively impact both what we understand and what we choose. The choice process interface represents only a starting point in terms of the observational enrichments required to further our understanding of these effects.

It is in some ways surprising that economists have focused so little prior attention on how well understood are the various options in any given decision-making context. While research has now begun on the many settings in which subjective "consideration sets" may be strictly smaller than the objectively available set of choices, nothing of equal power has replaced the principle of revealed preference.[5] It is the organizing power that this principle introduces that led to our experimental investigation of artificially enhanced choice data.

We see our agenda as illustrating one of the advantages of economic experiments over field experiments. Our experiments really require a controlled environment in which the process of choice is subject to "unnatural" manipulation and to observation. It would be hard, if not impossible, either to manipulate or to adequately observe the act of choice in a field experiment designed to be naturalistic.

NOTES

We thank Daniel Martin for his collaboration on the research reported herein, and we are grateful to Marina Agranov, Sen Geng, and Chloe Tergiman for their interest in and contributions to the broader research agenda. We thank Martin Dufwenberg and the

participants in the CESS Conference on Methods of Experimental Economics for valuable suggestions.

1. For each of the three choice set sizes, we generated 12 sets of values, which were used to generate the choice objects at both the low and the high complexity levels.

2. A graph of this distribution was shown in the experimental instructions. The beta distribution was chosen in order to "front load" the probability of a time being selected in the first minute of the choice round, as most subjects made their choices inside 120 seconds.

3. In support of this interpretation, many subjects indicated in a follow-up survey that they always selected their most preferred option.

4. One difference was that the pure choice experiments were run without time limits. When comparing with the choice process outcomes, CDM focus only on rounds from the choice process experiment in which the subject pressed the submit button before the allotted 120 seconds and thus did not hit the binding time constraint.

5. Rubinstein and Salant (2006) study choices made from sets presented in "list" order, effectively making the order of search observable. Masatlioglu and Nakajima (2013) characterize choices that result from iterative search of "consideration sets" related to each alternative. They focus on how final choice is related to an initial (externally observable) reference point.

REFERENCES

Armel, C., A. Beaumel, and A. Rangel. 2008. Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making* **3**(5):396–403.

Block, H. D. and J. Marschak. 1960. Random Orderings and Stochastic Theories of Response. In *Contributions to Probability and Statistics*, ed. I. Olkin. Stanford, CA: Stanford University Press.

Bronars, S. 1987. The Power of Nonparametric Tests of Preference Maximization. *Econometrica* **55**(3):693–698.

Camerer, C. 2008. The Case for Mindful Economics. In *The Foundations of Positive and Normative Economics: A Handbook*, eds. A. Caplin and A. Schotter. New York: Oxford University Press.

Camerer, C., G. Loewenstein, and D. Prelec. 2005. Neuroeconomics: How Neuroscience Can Inform Economics. *Journal of Economic Literature* **43**:9–64.

Campbell, D. 1978. Realization of Choice Functions. *Econometrica* **46**:171–180.

Caplin, A. 2008. Economic Theory and Psychological Data: Bridging the Divide. In *The Foundations of Positive and Normative Economics: A Handbook*, eds. A. Caplin and A. Schotter. New York: Oxford University Press.

Caplin, A. and M. Dean. 2008. Dopamine, Reward Prediction Error, and Economics. *Quarterly Journal of Economics* **123**(2):663–701

Caplin, A. and M. Dean. 2011. Search, Choice, and Revealed Preference. *Theoretical Economics* **6**(1):19–48.

Caplin, A., M. Dean, P. Glimcher, and R. Rutledge. 2010. Measuring Beliefs and Rewards: A Neuroeconomic Approach. *Quarterly Journal of Economics* **125**(3):923–960.

Caplin, A., M. Dean, and D. Martin. 2011. Search and Satisficing. *American Economic Review* **101**(7):2899–2922.

Ericsson, K. and H. Simon. 1984. *Protocol Analysis*. Cambridge, MA: MIT Press.
Gabaix, X., D. Laibson, G. Moloche, and S. Weinberg. 2006. Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *American Economic Review* **96**(4):1043–1068.
Gul, F. and W. Pesendorfer. 2008. The Case for Mindless Economics. In *The Foundations of Positive and Normative Economics: A Handbook*, eds. A. Caplin and A. Schotter. New York: Oxford University Press.
Ho, T. H., C. Camerer, and K. Weigelt. 1998. Iterated Dominance and Iterated Best Response in Experimental p-Beauty Contests. *American Economic Review* **88**:947–969.
Houtman, M. and J. A. H. Maks. 1985. Determining all Maximal Data Subsets Consistent with Revealed Preference. *Kwantitatieve Methoden* **19**:89–104.
Masatlioglu, Y. and Nakajima, D. (2013), Choice by iterative search. *Theoretical Economics* **8**: 701–728.
Payne, J. W., J. R. Bettman, and E. J. Johnson. 1993. *The Adaptative Decision Maker*. Cambridge, MA: Cambridge University Press.
Reutskaja, E., R. Nagel, C. Camerer, and A. Rangel. 2011. Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study. *American Economic Review* **101**(2):900–926.
Rubinstein, A. and Y. Salant. 2006. A Model of Choice from Lists. *Theoretical Economics* **1**(1):3–17.
Samuelson, P. 1938. A Note on the Pure Theory of Consumer's Behavior. *Economica* **5**:61–71.
Simon, H. 1955. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics* **69**(1):99–118.
Wang, J., M. Spezio, and C. Camerer. 2010. Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender–Receiver Games. *American Economic Review* **100**(3):984–1007.
Wilcox, N. T. 1993. Lottery Choice: Incentives, Complexity and Decision Time. *The Economic Journal* **103**:1397–1417.

CHAPTER 5

INTELLIGENT DESIGN: THE RELATIONSHIP BETWEEN ECONOMIC THEORY AND EXPERIMENTS: TREATMENT-DRIVEN EXPERIMENTS

MURIEL NIEDERLE

Introduction

When I interact with colleagues and friends who are venturing into experimental economics, as they either prepare for their own experiment or aim for a critical view of experiments run by others, I often hear the question: "What is a good experiment?" My first reaction is to answer my empirically minded colleagues "Well, let me ask you: What is a good regression?" Clearly a good experiment (or regression) is one that allows testing for the main effect while controlling for other plausible alternatives. This helps ensure that the original hypothesis is reached for the right reasons and the initial theory is not wrongly confirmed.

However, there is an aspect of experimental design that is probably closer connected to theory than empirical work: As designers, we are responsible for the environment in which the data are generated. As such, a good experimental design also needs to fulfill requirements one may impose on good theory papers: The environment should be such that it is easy to see what drives the result, and as simple as possible to make the point. The design has to provide a good environment for studying the questions at hand. Furthermore, ideally, the design (just like good theory) should be such that it seems plausible that the results could serve as prediction for behavior in other environments.

The design of experiments has some interplay with empirical methods because a good experimental design foreshadows how the data to be generated can be analyzed. As such, good design can often reduce the need for fancy econometrics, or, at times, allow econometrics to be more powerful. It also often implies that the experimental design has to take a stand on what it means to accept or reject an initial theory or hypothesis.

When deciding about a design, there are basically no fixed rules—the one exception probably being that economic experiments which use deception are really frowned upon, are hard to run in many experimental labs, and often have a hard time to be published in economics journals.[1] Apart from that, however, anything goes. This may make it harder to design and evaluate a good experiment.

In this chapter of the book on methodology of experiments, I want to focus on the interplay between experimental design and the testing of hypotheses. This includes hypotheses that rely on theory, as well as some that are described less formally. Specifically, I want to show how in many cases intelligent design can provide direct tests, instead of having to rely on indirect evidence.

The section entitled "Belief-Based Models—A Direct Test" shows a line of work where the test of the theory becomes more and more stringent, casting new doubts on the validity of the theory. In the section entitled "Experimental Design and Hypothesis Testing," I present several ways in which theory can be tested. First, I show how it might be useful to test assumptions directly rather than relying on econometrics. Then I discuss two methods to test whether a theory is responsible for the phenomenon. One can be called the "two-way" design and the other the "Elimination" design. Finally, I discuss issues that are relevant when running a horse race among theories. In the section entitled "What Channels Drive a Result? Treatment-Driven Experiments," I show how the methods used when testing theory apply even when there is no detailed model. I show how experiments can be used in trying to understand the important channels that drive a phenomenon.

When talking about theory and the relation to experiments, an unavoidable question will be: When should a theory judged to be "good" in terms of relating to the experimental results? What should the criteria be in the first place? I will touch on these issues as the chapter progresses.

Belief-Based Models—A Direct Test

The first example I want to delve into is how experiments have been used in the formulation and testing of new, less stringent theories on behavior in strategic games. I will present a series of approaches to test the theory, and I will show how new design made the test more and more direct and may lead to some new questions as to how to think of those theories.

It is well known that standard Bayesian Nash equilibrium makes very strong assumptions on rationality: First, players have to form beliefs about strategies of other players, and they have to best respond to these beliefs. Furthermore, in equilibrium these beliefs have to be correct. One simple modification is to relax the assumption that beliefs are correct while maintaining that players form beliefs and best respond to them. A prominent version of such a modification is the *k*-level thinking model.

This model was created and became prominent by the results of the so-called beauty contest or guessing game. In such a game, several players have to choose a number from some interval—for example, 0 to 100. Furthermore, the player who is closest to, say, 2/3 of the mean of all players receives a prize. Obviously the Nash equilibrium is 0, but 0 is also never the winning guess. In a first experiment, Nagel (1995) showed that a large fraction of responses center around 2/3 of 50 and around 2/3 of 2/3 of 50 (see Figure 5.1).

The behavior can be rationalized the following way. Those that play 2/3 of 50 may be participants that believe that other players simply pick numbers randomly, and hence the best response is to guess 2/3 of 50. Such players are called level 1 players (and correspond to step 1 in Figure 5.1), as they best respond to level 0 players, players that play nonstrategically, and here are assumed to select numbers

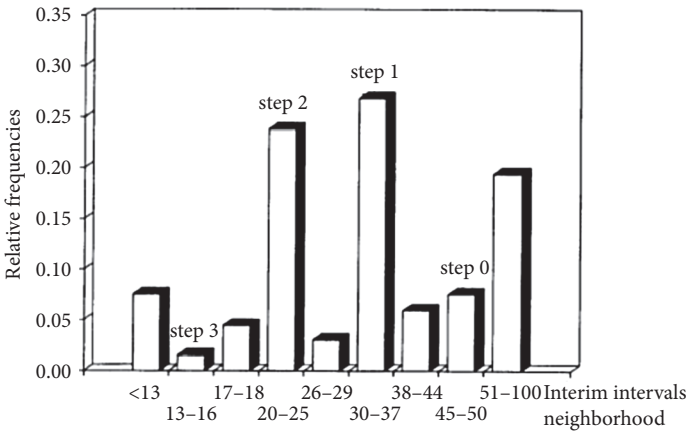


Figure 5.1. Relative frequencies of voices in the first period according to the *k*-level classification with a level 0 expected play of 0. *Source*: Nagel (1995).

randomly. Participants that choose 2/3 of 2/3 of 50 best respond to a world that is fully comprised of level 1 players, and are hence called level 2 players. In general, for any given strategy of level 0, a level k player plays a best response to the belief that all other players are of level $k - 1$ (see also Stahl and Wilson (1995)).[2]

Estimating Parameter(s)

Many experiments on belief based models, such as Nagel (1995), show that k-level models are pretty good at organizing behavior observed in the lab and in the field.[3] The data are analyzed by estimating which fraction of players use strategies that are close to k-level strategies for various k's and for an appropriately chosen level 0 play. Such participants are then classified as k-level players.[4] In principle, such estimations allow a lot of freedom in the choice of a level 0 strategy, which may make fitting data to k-level play easy. The literature has in general put a restriction on level 0 play as random play with all actions chosen with the same probability (see, e.g., the discussion in Camerer et al. (2004)).[5] In the case of common value auctions with private information, "naive" play has also been proposed as a natural level zero play. Naive play is often a strategy that depends in a naive way on the agents' private information; often they are simply assumed to bid their signal (see Crawford and Iriberri (2007)).

The experiments in this category do not allow us to more directly test whether the comparative static predictions of a level k model are correct, or, even more fundamentally, whether players actually form beliefs and best respond to them, which is the assumption at the heart of any belief-based model, as well as, of course, the standard model.

In general, experiments that estimate parameters only allow for very limited conclusions. This may not be a problem when the aim is to estimate a specific parameter, such as, say, coefficients of risk aversion, while maintaining the assumption that players have well formed preferences over risky outcomes. Also this may not be a limitation when finding a parameter that suggests a theory is wrong. It is, however, more of an issue when trying to conclude that a specific theory is right. This may be especially the case when, at the same time, there is no accompanying discussion about what fraction of potential behavior can be accommodated by the theory and about what fraction of potential behavior would reject the theory (where of course it may be hard to say what kind of behavior can reasonably be expected to occur).

Comparative Static Predictions

Moving beyond experiments that (merely) serve to estimate a parameter, comparative static experiments on k-level thinking provide a somewhat more stringent test than fitting the data of an experiment.

For example, Costa-Gomes and Crawford (2006) test whether participants play differently in generalized two-player guessing games, depending on the strategy used by the opponent. In such a game, each player receives a lower limit and an upper limit in which to choose a number and a personal multiplier. The amount the player is paid is dependent on how close their guess is to the number that equals their multiplier times their opponent's guess. They want to assess whether participants respond to strategies of opponents. There are two kinds of ways of doing that. One is to elicit a subjects' beliefs about the strategy used by the opponent. This has the problem that gathering information about beliefs is hard and that it may or may not influence actual play (see, e.g., Costa-Gomes and Weizsaecker (2008)). A second approach is to manipulate the beliefs of a subject about the strategy of their opponent. This is the route chosen by Costa-Gomes and Crawford (2006).

Specifically, they have each participant play against a fixed strategy played by a computer that is explained to them in detail. For example, if the computer is programmed to play level 1, where level 0 is random play, the description says (see the web appendix to Costa-Gomes and Crawford (2006), on http://dss.ucsd.edu/~vcrawfor/accessed January 14, 2010).

> The computer's rule is based on the assumption that you are equally likely to choose any guess from your lower limit to your upper limit, so that, on average, you guess halfway between your lower and upper limits. The computer's rule is to choose the guess that would earn it as many points as possible if you guess halfway between your lower and upper limits.

For level 2 the description is:

> The computer's rule is based on the assumption that you assume that the computer is equally likely to choose any guess from its lower limit to its upper limit, so that, on average, it guesses halfway between its lower and upper limits.
>
> The computer's rule is to choose the guess that would earn it as many points as possible if you chose the guess that would earn you as many points as possible assuming that the computer guesses halfway between its lower and upper limits.

They say that participants are indeed able to best respond, to some extent.

Georganas et al. (2009) also offer a more direct test of the hypothesis that participants change their level of play dependent on their opponent, where they manipulate the beliefs about the ability of the opponent. They have participants take an IQ test amidst several other tests, to construct a combined score. Participants played 10 games (generalized two-player guessing games and other games): once against a randomly drawn opponent, once against the opponent with the highest combined score, and once against the opponent with the lowest combined score. The idea is that participants should adjust the level of strategic reasoning they assign to their opponent, depending on the opponents' general abilities. While they find that participants use somewhat higher level-k strategies against the opponent with the highest score compared to a random opponent, there is no significant

downward shift in level when playing against the person with the lowest score. Furthermore, they find a lot of variation in a participant's level of sophistication across games. This is true not only at the absolute level, but also when comparing the relative sophistication of a subject compared to the depths of reasoning of all other participants.

Overall, comparative static tests of k-level thinking models seem to have limited success. Furthermore, across games, players cannot necessarily be described as being of a fixed level k. Similarly for the cognitive hierarchy models, estimations of the parameter of the Poisson distribution of level k's yields differences across games, though Camerer et al. (2004) summarize their paper as "an average of 1.5 steps fits data from many games." Two remaining open questions are, For what classes of games can a large fraction of "reasonable" behavior lead to results that are not consistent with the model, and what games do allow for the k-level model to be falsified? A different issue is how to think of the phrase "many games," one I will return to in the next section.

Are the Underlying Assumptions Really Fulfilled?

The most stringent interpretation of many papers on k-level models is that they indeed present a fair representation of the way in which many participants behave. This means that participants form beliefs that other players are of some level k and best respond to them by playing a level $k + 1$ strategy. The tests so far involve estimating whether a sizeable fraction of plays, and/or of players correspond to strategies that fall into the level k description (for an appropriately chosen level 0). The more stringent test of a comparative static prediction that players adapt their play given the strategies of others has more mixed success.

How can we test directly whether players actually form beliefs and best respond to them? The strategy of manipulating beliefs instead of trying to assess them seems right. However, when using the approach of Costa-Gomes and Crawford (2006), one might worry that by describing, say, the level 1 strategy of the computer, participants may be "trained" in thinking about best responding, and hence may be more likely to become a level 2 player. Providing players with a description that includes best response behavior may change their own thought process. On the other hand, the strategy of Georganas et al. (2009) rests upon players forming beliefs what strategies players use, when they are either the winner or loser of a quiz. While intelligence may be correlated with the depth of reasoning a player may be capable of, it need not be correlated with their beliefs about the level of reasoning used by their opponent. As such, clear predictions may not be that straightforward.

Ideally, we would like to know the beliefs that players have about their opponents' strategy, without influencing their thought process, explaining strategies, or eliciting additional information. That is, we would like to provide a direct test whether players best respond to beliefs they have about strategies of their opponents. We can achieve this by designing an environment where both the players and

the experimental economist know the strategy of the players opponent, without providing the player with any additional information.

The environment we use to directly test k-level thinking models is overbidding in common value auctions, where participants in general fall prey to the winners curse. Overbidding, bidding above the Bayesian Nash equilibrium, has been shown to be consistent with k-level thinking, and indeed is one of their success stories (see Crawford and Iriberri (2007)).[6]

In Ivanov et al. (2010), we propose a very simple two-player common-value auction. Players each receive a signal x between 0 and 10 (discrete values only), where the value of the item is the maximum of the two signals. Participants then bid in a second price-sealed bid auction for the item. It is easy to see that bidding less than one's signal is a weakly dominated strategy, as the item is always worth at least one's own signal. Furthermore, a second round of iteration of weakly dominated strategies eliminates all bid functions that call for bids that are strictly greater than one's signal. In fact, bidding the signal is the unique symmetric Bayesian Nash equilibrium bid function.

The general finding in common-value auctions is the winners' curse; that is, participants bid above the Bayesian Nash equilibrium, which often results in the winner of the auction paying more than the value of the item. How can the k-level thinking model in this simple environment generate the winners' curse? Start with a level 0 player who bids randomly. Then the best response entails overbidding. The intuition is that in this case a level 1 player bids against a random number in a second price auction. Hence, the best response to random bids is to bid the expected value of the item, which is, in general, strictly higher than the received signal, apart from the case when the signal is 10. Therefore, a level 1 player would be overbidding, compared to the symmetric Bayesian Nash equilibrium. This implies that when two level 1 players bid against each other, we would confirm the standard result of a winners curse.

The nice feature of our setup is that a best response to a level 1 bidder (or any bidder who bids above their signal) is to bid the signal. In general, a best response to overbidding entails to certainly bid less than that. This allows for a very stark prediction of any model that includes best response behavior. Hence we have a very simple comparative static prediction if we manipulate the beliefs of at least some players, that their opponent is someone who bids above their signal. However, the goal is to achieve this without providing players with any information about possible strategies, so as not to affect their thought process.

In the experiment, participants first play for 11 rounds the two-player common-value second-price sealed-bid auction against varying opponents, where each participant receives each signal $\{0, 1, 2, \ldots, 10\}$ exactly once. As such, we basically elicit the participants' bid function. As is common in the literature, to not distort the information that participants have about the game, they receive no feedback and no information about whether they won the auction, or what the item was actually worth.

Figure 5.2 shows for each signal the fraction of bids that fall into various categories, where we allow for small errors and hence have bands of 0.25 around the signal. Note that bids $b(x) < x - 0.25$ and $b(x) > 10.25$ can only be level 0 bids, bids $x + 0.25 < b(x) \leq 10.25$ can be classified as level 1 bids (for a random level 0), and bids $b(x) \sim x$ can be classified as level 2 bids.

Note that for each signal x, all bids that are strictly below x are weakly dominated since the item is always worth at least x. Similarly, bids strictly above 10 are weakly dominated. Figure 5.2 shows that for each signal, the majority of bids can be classified as level 1 bids for a random level 0 player, because those participants place bids above their signal but not above 10. A sizeable fraction of bids correspond to the symmetric Bayesian Nash equilibrium and can be thought of as level 2 bids.

Table 5.1 classifies bidders depending on where they place the majority of their bids (6 out of 11).[7] In Phase I, many bidders place the majority of their bids in a way

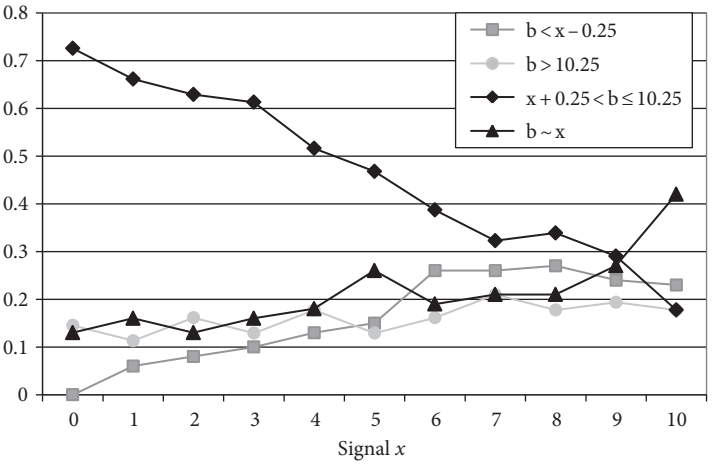


Figure 5.2. For each signal, the fraction of bids that fall into various categories: "$b < x - 0.25$" represents bids such that for a signal x the bid $b(x) < x - 0.25$. Similarly for "$b > 10.25$" and "$x + 0.25 < b \leq 10.25$." "$b \sim x$" includes for each signal x bids in $x - 0.25 \leq b \leq x + 0.25$.

Table 5.1. Subject Classification in Phase I of the Baseline Treatment

	Phase I
Underbidders	5
Signal bidders	9
Overbidders	25
10+: Above-10 bidders	10
Ind.: Indeterminate	13

consistent with *k*-level thinking. While the 10 bidders who bid above 10, as well as the 5 underbidders, are clearly not rationalizable with *k*-level thinking, a total of 25 (40%) participants can be classified as level 1 bidders while 9 (15%) can be classified level 2 (or higher) bidders.

So far, the experimental results provide another "success" story of *k*-level thinking. The majority of bidders and bids can be classified as level 1 or level 2. However, we now turn to a more direct test of the *k*-level model.

In the main treatment, Phase II of the experiment, participants play again 11 rounds of the two-player common-value second price-sealed bid auction, where once more they receive each signal exactly once. This time, however, they play against a computer. The computer, upon receiving a signal y, will place the same bid that the participant placed in Phase I when receiving that signal y. That is, the computer uses the participants' own prior Phase I bid function. Hence, in Phase II, participants bid against their old bid function. This method allows us to have perfect control of a subjects' beliefs about the other players' strategy, since it is simply the subject's own past strategy. Furthermore, we achieved this without any interference, without providing any new information that may change the mental model of subjects. While the data of the treatment I present does not remind participants of their old bid function, we do so in another treatment, with virtually identical results.

Any best response model, and as such also the *k*-level model, predicts that players best respond to their own past strategy. Hence, a player of level k is expected to turn into a level $k + 1$ player. This implies that overbidding should be reduced in Phase II compared to Phase I. Consider a participant who overbids in Phase I. Then, bidding the signal is a best response. Continuing to overbid, but less so, may or may not be a best response, depending on how much the participant overbid beforehand and by how much the bid function is lowered. However, no change in behavior is clearly not a best response.

Figure 5.3 shows for each signal the fraction of bids that fall into level 1 play, and those that are close to bidding the symmetric equilibrium, both for Phase I (filled) and Phase II (hollow). The fraction of bids, both above and around the Bayesian Nash equilibrium, are virtually unchanged.

Furthermore, we characterize bidders depending on where they placed the majority of their bids, and we determine whether bidders of various types changed their classification.

Table 5.2 shows that overbidders mostly remain overbidders (56%) and that only a minority (24%) turn into underbidders or signal bidders. For subjects who are overbidders in parts I and II, we find that only 23% of bids in Phase II are best responses to Phase I behavior. By not behaving optimally in part II, these subjects are, on average, foregoing more than 20% of the earnings an average subject made in the course of the experiment.

We investigated whether the winners' curse in common value auctions can be rationalized using belief-based models such as *k*-level thinking. We achieved

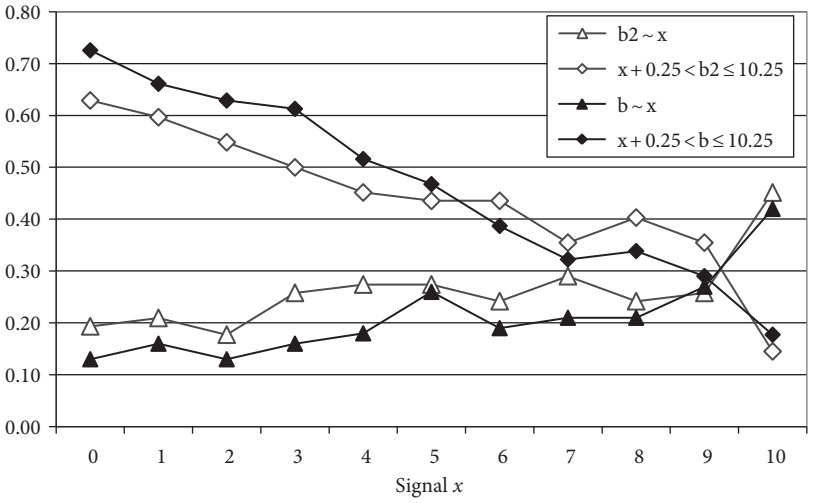


Figure 5.3. For each signal, the fraction of bids that fall into various categories: "$x + 0.25 < b \leq 10.25$" represents bids such that for a signal x the bid $x + 0.25 < b(x) \leq 10.25$. "$b \sim x$" includes for each signal x bids in $x - 0.25 \leq b \leq x + 0.25$. b indicates a bid in Phase I and $b2$ a bid in Phase II.

Table 5.2. Subject Classification in Parts I and II of the Baseline Treatment, Depending on How They Placed the Majority (6 out of 11) of Their Bids

	Under	Signal	Over	10+	Ind.	**Phase I**
Underbidders	2	0	2	1	0	**5**
Signal bidders	0	5	3	1	0	**9**
Overbidders	1	5	14	1	4	**25**
10+: Above-10 bidders	2	1	1	6	0	**10**
Ind.: Indeterminate	2	2	3	5	1	**13**
Phase II	**7**	**13**	**23**	**14**	**5**	

a direct test of this hypothesis by comparing behavior in environments in which overbidding can be rationalized (such as in Phase I play) and in environments where it cannot (such as in Phase II play for participants who were overbidders in Phase I). The results of the experiment are, in general, bad news for any theory that relies on best response behavior, especially in complicated environments such as second-price sealed-bid common-value auctions. This is despite the fact that our environment seems much easier than some other environments used in common-value auction experiments.

This poses questions as to what we think a theory should accomplish. I will restrict attention here to musing about the role of theory in understanding behavior

in experiments. Certainly, theory has a big role to play when, for example, deciding to give advice, whispering into the ears of princes, and performing real design (see Roth (2008)). Should a theory be merely a tale, providing insights, but otherwise, like a fable, clearly wrong, clearly leading to absurd conclusions or predictions, but, once more like a fable, with a kernel of truth? This view is certainly not unheard of (see Rubinstein (2006)).

Alternatively, one might want a theory to fit data very well, and so a theory might be judged by how well it reconciles with a large set of data. This is the approach taken by many proponents of new belief-based models. These models fit the data of the experiments better than standard theory, and as such they have been deemed superior in providing a model of how agents behave in such environments.

Finally, one might want a theory to provide good (i.e., accurate) predictions, either in general or at least in terms of comparative statics, to guide us in predicting how behavior would change, if we were to make certain changes to a game. As such, a good theory may also provide insights as to what are the important parameters of a game that are likely to affect behavior in the first place.

Note that, in principle, one theory might be a better able to fit data on any given experiment, while another theory might be better at making predictions out of sample.

In this section, the model of k-level thinking seems to be less promising as a model that is absolutely right, which makes good comparative static predictions when changing the game. Though maybe, in this game, participants are so much at a loss that the model simply does not apply, that maybe, using it to explain behavior in common value auctions was simply too ambitious a goal? It still remains a question of whether the model might be able to predict behavior in other instances. As such, this goes back to the question of whether we should be content when a model can fit the data of multiple experiments (though maybe with varying parameters, or varying proportions of level 1 and level 2 players). Or do we require the model to make predictions out of sample? This can include a good fit on new games, which is the approach that has mostly been taken by level-k proponents so far. It can also include predicting comparative static behavior better than other models (which has received less attention so far).

Finally, assume the model cannot predict out-of-sample behavior, by trying to fit behavior that is obtained when certain parameters of the game change (such as manipulating the beliefs about the strategy of the opponent). It might still be the case that the description of the data in itself might be valuable. For instance, it could be that the fraction of level 2 players (or the specific k that fits the cognitive hierarchy model to the data of a specific game) may be a good predictor of, for example, how "difficult" a game is. It could be a decent proxy for the complexity of a game that could be used to further our understanding on what it is that makes some games harder than others.

Later in the next section, I will revisit the question of how to think of theories in the face of experimental evidence.

Experimental Design and Hypothesis Testing

The first example showed how design can allow for a direct test, instead of simply having to indirectly estimate whether behavior confirms to the model at hand. Note that the design reduced the need for econometrics. In this chapter, I want to provide a few more examples of that sort, where intelligent design changed the problem so that the hypothesis could be attacked directly. However, I also want to show how sometimes intelligent design is needed in order to be able to provide an environment in which a hypothesis can be tested in the first place. The following can be thought of as a toolbox for designing experiments.

Testing for Selection

In this first section I want to elaborate on a design that allows for a comparison between the power of an intelligent design and the result of using standard econometric techniques. Once more the environment is one of overbidding in common value auctions, this time first price auctions.

While overbidding (bidding above the risk-neutral Bayesian Nash equilibrium) is the standard behavior in common-value auctions, later rounds of experiments show fewer instances of participants falling prey to the winners' curse compared to earlier rounds. Furthermore, experienced participants, participants who return sometimes weeks later, suffer from the winners' curse less than those who participate for the first time. There are two potential reasons for why participants seem to learn to use better strategies. One is that less able bidders may simply go bankrupt in the course of the experiment (having suffered repeatedly from the winners' curse) and are barred from further participation, as they have no more earnings they can lose. Furthermore, when experience is measured across sessions taking place on separate days, participants who overbid a lot and made less money may not return at the same rate for subsequent experimental sessions. Hence, one reason for better performances in later rounds could be a pure selection effect. Second, it could be that bidders indeed learn to avoid the winner's curse, where learning can occur in the first session, between sessions, and during the second session.

One aim of Casari et al. (2007) is to design an experiment that can directly measure the effects of selection, compared to learning. To reduce the effect of selection, they have treatments in which some participants have higher initial cash balances than others and throughout the experiment also receive windfall gains

(via a lottery), thereby affecting the probability with which a subject will go bankrupt during the course of the experiment. To study the importance of selection in accounting for the fact that experienced participants perform better, Casari et al. (2007) in their two-week experiment vary the show-up fees for the second week, and for some participants even hold half the earnings of the first week in escrow, to ensure a high return rate for a subset of participants.[8]

The control treatment employs standard procedures: Participants receive a show-up fee of $5 and an initial cash balance of $10. All subjects were invited back to week 2, where they received the same show-up fee and cash balance once more. In the bonus treatment, starting cash balances were $10 for half the participants and $15 for the other half. Furthermore, after each auction, each active bidder participated in a lottery that paid $0.5 with 50%. In addition, a show-up fee of $20 was paid only after completing week 2's session, with 50% of the earnings of the first week held in escrow as well. A third treatment was similar to the bonus treatment except that participants all received a show-up fee of $5 in week 1 and either $5 or $15 in week 2.

Within the first session, the changes in design indeed affect bankruptcy rates, which range from 46.3% to 20.3%. Similarly, the rates of participants returning for a second experimental session vary from 96% to 60%. A first clear indication of selection of participants is to consider, among the subjects that went bankrupt in the first session, how many return to a second session. In the baseline treatment, this is only 47.7%, while it is 88% in the treatments that made bankruptcy harder in the first place! Nonetheless, in all treatments the authors find that participants learn that their behavior is closer to the RNNE, albeit to various degrees. It seems that both market selection and individual learning is responsible for improved outcomes of experienced bidders.

The authors also show that standard econometric techniques fail to provide evidence of unobserved heterogeneity in the bidding behavior of participants and fail to detect any selection effects.

The paper shows how intelligent design can address selection (or potentially other econometric issues) directly, instead of having to rely on sophisticated econometric techniques, which may fail to find any evidence (which could be due to the fact that the samples of experimental economists are typically smaller than those of labor economists).

Two-Way Design

Overbidding—that is, bidding above the risk-neutral Nash equilibrium (RNNE) by participants—is not only a regular phenomenon for common-value auctions, but also in private-value first-price auctions. Early work has attributed this to risk aversion (see, e.g., Cox et al. (1988), which calls for bids above the RNNE, closer to one's valuation. The literature has spawned a serious critique, namely that behavior of participants in first-price auctions may be hard to take seriously, as the

change in expected payoff is quite flat around bid functions that correspond to the risk-neutral Nash equilibrium (see Harrison (1989)). This spawned a very lively debate.[9]

In a very clever design, Kagel and Levin (1993) want to show that, in general, overbidding in first price auction may not necessarily be attributable to either risk aversion or the possibility that participants simply don't care about the bids they place.

One way to address this question would be to try to estimate each participant's level of risk aversion and correlate it with the bids they place. Note that we would still need to heavily rely on the Nash equilibrium model, implying that subjects form correct beliefs about the degrees of risk aversion of other participants, and so on.

Kagel and Levin (1993) present a much more direct, simpler, and elegant solution. Specifically, they have participants bid not only in first price but also in third price auctions, auctions in which the highest bidder receives the object for the third price. In this case, the RNNE calls for bidding *above* one's valuation, and risk aversion calls for bids *below* the RNNE. Both of these are in contrast to behavior expected in the first price auction, where RNNE calls for bids below one's valuation whereas risk aversion calls for bids above the RNNE.

They find that bidders bid above the RNNE in both auctions formats, which makes it more unlikely that the overbidding in the first price auction can be attributed to risk aversion (see also Kagel (1995) and Kagel and Levin (forthcoming)).

In this subsection we showed how an intelligent design can cast doubts on a prominent theory for a phenomenon. This was achieved by changing the environment such that the prominent theory (risk aversion) would now make opposite predictions to other theories that may account for the initial phenomenon, such as "wanting to win" for the "bidding above the risk-neutral Nash equilibrium in first price auctions" phenomenon.

To elaborate on the use of such a two-way design, I want to provide maybe one of the earliest examples of such a design from the book of Judges of the Old Testament, Chapter 6 (by courtesy of Al Roth's experimental economics class).

The story begins with the Israelites, turned away from God after 40 years of peace brought by Deborah's victory over Canaan, being attacked by the neighboring Medeanites. God chose Gideon, a young man from an otherwise unremarkable clan from the tribe of Manasseh, to free the people of Israel and to condemn their worship of idols. Gideon, to convince himself that the voice he hears is indeed the voice of God, asks for a test:

> And Gideon said to God: "If You will save Israel by my hand, as You have said, look, I will put a fleece of wool on the threshing-floor; if there be dew on the fleece only, and it be dry upon all the ground, then shall I know that You will save Israel by my hand, as You have said."
>
> And it was so; for he rose up early on the next day, and pressed the fleece together, and wrung dew out of the fleece, a bowlful of water.

So far, this makes Gideon merely an empirically minded person, not a good designer of experiments. Gideon, however, realized that there could be an alternative explanation for the main result. It could be that this is simply the way it is; after all, he probably wasn't used to leaving a fleece outside. His design so far cannot distinguish whether the result is due to God almighty, or merely the way cold nights interact with a fleece left outside. In his next design, he removes the "natural" explanation and he tests whether it is God's doing. Specifically, he asks whether God can reverse the result that may be due to nature only.

> And Gideon said to God: "Do not be angry with me, and I will speak just this once: Let me try just once more, I ask You, with the fleece; *let it now be dry only upon the fleece, and upon all the ground let there be dew.*"
>
> And God did so that night; for it was dry upon the fleece only, and there was dew on all the ground.

The experiment also allows for predictions outside of the experimental design: Gideon raised the army which indeed defeated the Medeanites.

This is a prime example of an intelligent design: Change the environment such that the hypothesis of choice (God) would reverse the result, while the alternative (Nature) would leave the outcome unchanged. While many designs do not necessarily have comparative statics that work that beautifully, these are often useful in convincing the audience that the results are indeed driven by the specific hypothesis at hand.

Elimination Design: Testing a Theory by Eliminating Its Applicability

Apart from the two-way design, there is another prominent method one could call the "Elimination" design, to cast doubt on the applicability of a theory to a particular phenomenon. Change the environment in a way that the problem mostly stays the same while, however, eliminating the conditions that allow the theory at hand to account for the phenomenon. That is, instead of testing the theory directly, examine to what extent similar behavior can be found in environments in which the model has no bite. If behavior remains unchanged, at least it implies that other factors may be at work as well. I will present two examples in detail that make that point, both of which are in environments which I already presented.

The first example concerns the guessing game which was introduced in the section entitled "Belief-Based Models—A Direct Test," which was a breeding ground for new theory, such as *k*-level thinking. The next experiment takes a phenomenon, such as players guessing a number not equal to zero in a guessing game, and changes the environment in a way to eliminate rationalizations of such behavior due to *k*-level models.

Grosskopf and Nagel (2008) change the guessing game to have only two, instead of three or more players. The rule is that the person who is closest to two-thirds of the average wins. With two players this translates to the winner being the one who chooses the smaller number. That is with two players, guessing 0 is a dominant strategy. Hence any model that relies on players having nonequilibrium beliefs about the opponents to justify non-equilibrium behavior has no bite, as there exists a unique dominant strategy.

Grosskopf and Nagel (2008) find that among students, the guesses are virtually identical to the guesses made when the number of players was larger than 2, specifically, the instances of 0 were the same about 10% in both cases. While professionals (game theorists) are more likely to choose 0 when there are two rather than three or more players, zero is still chosen only by about 37%.

The findings of Grosskopf and Nagel (2008) cast serious doubt on the need for *k*-level models to explain guesses above 0, since such guesses are common even when 0 is a dominant strategy. However, there remains the possibility, that, when there are dominant strategies, other biases become important, or that participants were simply confused and did not take into account that there were only two players.

A second example of an elimination design I want to provide is given in Ivanov et al. (2010). Remember that one way to justify overbidding—that is, bidding above the signal in the simple two-player second-price sealed-bid common-value auction—is to assume that the opponent sometimes bids below the signal. In one treatment, we eliminate the possibility for participants to place a bid that is strictly below the signal (we call it the MinBid treatment).

We can compare the proportion of bids, and the proportion of bidders who can be classified as bidding above the signal in Phase I of the Baseline treatment (the treatment which was discussed at length in the section entitled "Are the Underlying Assumptions Really Fulfilled?"), where participants play against a random person and could place any bids they wanted below 1,000,000. If the main reason for overbidding in the Baseline treatment is that players believe that others may be playing randomly and sometimes underbid, then we would expect a large reduction of overbids in the MinBid treatment. However, we find the opposite, an increase in the fraction of overbids, from about 40% to 60%. Overbidding is probably more frequent in the MinBid treatment because underbidding is impossible so that all bids are distributed in three, rather than four, categories. Given this, the frequencies of overbidding seem quite comparable.

The findings of Ivanov et al. (2010) cast doubt on belief-based models being the driving factor behind overbidding in common-value auctions. While the third treatment, the MinBid treatment, eliminates any explanatory power of the theory, it could still be, potentially, that other biases that are of a similar magnitude become important.

To summarize, one way to weaken the hypothesis that a theory can account for a phenomenon is to remove the applicability of the theory and show that the

underlying phenomenon is virtually unchanged. However, it is still potentially possible that in this slightly changed environment there are other forces at work that yield similar results. It does not directly exclude that the theory has (at least also) some explanatory power for the phenomenon at hand. As such, providing a direct test may prove more fruitful to convince proponents of the theory than would such indirect, albeit quite elegant, tests.

Running a Horse Race Among Theories

Finally, I want to come to a quite popular method when comparing the predictive power of different theories. Before, I argued that theories may be valuable both in making point predictions or, alternatively, in predicting comparative statics due to changes in the environment. When it comes to papers that run a horse race among theories, the valuation is, in general, driven by how well each theory is able to fit the data in a fixed set of games; almost no attention is given to comparative static predictions.

In many papers such horse races involve a preferred theory (mostly a new theory of the authors of the study) and some more "standard" or established theories. Often, authors would pick a few games, have subjects play those games, and then compare the results across these games. It is not uncommon to run regressions often showing how a certain favorite theory does better than other theories, when giving equal weight to each of the games selected.

Here are my two biggest concerns with papers of that sort: The first is a lack of a deep discussion of how the games have been chosen. Often, such discussion is short and not precise. This makes it hard to interpret any econometrics based on those games: What does it mean that the author can pick, say, 10 games in which the authors' model does better than the other models? What does it even mean to run such a regression? Clearly, few people would be impressed if the authors tried many more games and then selected a few such that the authors' preferred theory wins the horse race. While such blatant abuse is probably rare, often authors may simply more easily think of games that confirm their intuition.[10] It may be hard to assume that the choice of games is not in some way biased. Of course, a potentially equally big problem is if the winning theory is designed after the fact, as I hope to make clear in the next paragraph.

This critique goes often hand in hand with my second problem with such work: One has to be careful that new theories are not what could be called "toothbrush" theories—that is, theories that one wants to use only on one's own data (just like a toothbrush is for a single user only). What it means is that many papers take the form: "*My* theory works better on *my* data than *your* theory works on *my* data." As such, it is clear why that is often not that impressive. . . .

A paper that runs a horse race among theories has to first decide what success means. Is success really better predicting behavior in a few very specific games? Or is success predicting behavior in a given class of games?

In this chapter I want to advocate for the latter, which is nicely described in Erev et al. (2007). They test the predictive power of various models in two-player zero-sum games, including, among others, standard Nash predictions and learning models. In order to do that, they have participants play in a set of zero sum games with a unique mixed strategy Nash equilibrium. However, since the models are supposed to make a good prediction on that whole class of games, the authors did not pick the games themselves, but rather chose them randomly. This is a good strategy, unless there is a very good reason to focus only on a specific set of games.

That it may be easy to even fool oneself (taking the view that authors hopefully did not want to fool only others) when choosing games can be seen in the literature that precedes Erev et al. (2007) and Erev and Roth (1998). Previous papers can roughly be put in two groups: papers that concluded that Nash is a good description of behavior, and papers that did not. It seems, however, these two groups can also be characterized the following way: Proponents of Nash tended to pick two-player zero-sum games where the mixed strategy equilibrium resulted in relative equal payoffs for both players. The others tended to pick games where this was not the case. As no paper actually made that point precise, I imagine it was not due to earlier authors trying to fool others; their intuition may simply have led them to pick games that confirm their initial hypothesis.

Most recently, there have been several contests whose goal was to find a model that predicts well in a class of games, where the exact games to be played will be randomly drawn (see Erev et al. (2010a, b)).

What Channels Drive a Result? Treatment-Driven Experiments

So far I have focused on experiments that test theories with a very precise underlying model. The experiments in this section go mostly beyond simple parameter testing or testing the comparative static of a well-developed theory that makes precise predictions. The aim here is to go deeper into understanding the mechanism behind an initial finding. What are the channels that drive the initial result? Note that devising treatments to better understand the channels behind a specific result can of course be done whether the initial finding is well grounded in theory or not. Similarly, the possible channels may be given by economic theory, though they may also be derived by findings from other disciplines, such as psychology.

The focus of this chapter is to describe the power of experiments in terms of understanding the channels that drive a result. This allows us to turn on and off various channels, which allows us to measure their impact and really understand what drives a phenomenon. It is this powerful ability to control the environment that makes experiments so useful, especially when trying to understand what mechanism is responsible for the initial result, which are the key channels.

In this section, I will focus on the topic of whether women shy away from competition and possible explaining factors. While Niederle and Vesterlund (2007) do not test a specific theory, we can ask, theoretically, What are potential economic reasons for (gender) differences in the decision to enter a tournament? Since we use a within subjects design, we need to have all controls ready before we start running the experiment (which is what makes such designs harder to implement). The advantage is that it will be easier to discern how much various factors contribute to the decision to enter a tournament, as opposed to merely being able to say that they have some impact.

We aim to test whether women and men enter tournaments at the same rate, where their outside option is to perform in a noncompetitive environment. Clearly, one reason for gender differences in choices of compensation scheme can be gender differences in performance. Gneezy et al. (2003) show that average performances of women and men in tournaments may be very different from each other, even when performances under a piece rate incentive scheme were very similar. In fact, the paper shows a significant change in gender differences in performance across incentive schemes. Hence, in NV we aim for a task in which performance does not vary too much with the incentive scheme. Furthermore, all performances, both in competitive and noncompetitive environments, are assessed when determining the choices of women and men. Beyond that, what are possible contributing factors to gender differences in choice of incentive scheme?

Explanation 1: Men enter the tournament more than women because they like to compete.

This will be the main hypothesis, so we have to design the experiment such that we can rule out other explanations.

Explanation 2: Men enter the tournament more than women because they are more overconfident. Psychologists and economists often find that while both men and women are overconfident about their relative performance, men tend to be more overconfident than women (e.g., Lichtenstein et al. (1982), Beyer (1990), Beyer and Bowden (1997) and Mobius et al. (2010)).

Explanation 3: Men enter the tournament more than women because they are less risk averse. Since tournaments involve uncertain payoffs, potential gender differences in risk attitudes may affect the choice of compensation scheme.[11]

Explanation 4: Men enter the tournament more than women because they are less averse to feedback. One consequence of entering the tournament is that the individual will receive feedback on relative performance.[12]

The experiment has groups of 2 women and 2 men seated in rows, and we point out that participants were grouped with the other people in their row. While participants could see each other, we never discuss gender during

the experiment.[13] The task of our experiment is to add up sets of five two-digit numbers for five minutes, where the score is the number of correct answers. After each problem, participants learn the number of correct and wrong answers so far, and whether the last answer was correct. Participants do not receive any feedback about relative performance (e.g., whether they won a tournament) until the end of the experiment.

The experiment has four tasks, where one of which will be randomly chosen for payment at the end.

Task 1—Piece Rate: Participants are given the five-minute addition task. If Task 1 is randomly selected for payment, they receive 50 cents per correct answer.

Task 2—Tournament: Participants are given the five-minute addition task. If Task 2 is randomly selected for payment, the participant who solves the largest number of correct problems in the group receives $2 per correct answer while the other participants receive no payment (in case of ties the winner is chosen randomly among the high scorers).[14]

In the third task, participants once again perform the five-minute addition task, but this time they select which of the two compensation schemes they want to apply to their future performance, a piece rate or a tournament. The choice between the piece rate and the tournament should allow predicting money maximizing choices. Hence the choice must be independent of the subjects' beliefs about other players' choices which would otherwise enter the maximization problem and hence make it theoretically difficult to make money-maximizing predictions. This implies that the choice of each participant cannot influence other participants' payoffs.

Task 3—Choice: A participant who chooses piece rate receives 50 cents for each correctly solved problem. A participant who chooses tournament has the new task 3 performance compared to the previous task 2 tournament performance of the other participants in his or her group. If the participant has the highest performance, she or he receives $2 for each correct answer, otherwise she or he receives no payment. This way a choice of tournament implies that a participants' performance will be compared to the performance of other participants in a tournament.

Furthermore, since a participant's choice does not affect the payment of any other participant, we can rule out the possibility that women may shy away from competition because by winning the tournament they impose a negative externality on others.[15]

Task 4—Choice of Compensation Scheme for Past Piece-Rate Performance: Participants decide between the piece rate and tournament incentive scheme for their task 1 piece rate performance, where a tournament choice results in a payment only if the participant had the highest task 1 piece rate

> performance in their group. This choice mimics the task 3 choice while eliminating any tournament performance. Specifically, this choice, like the choice in task 3, requires that participants decide between a certain payment scheme (piece rate) and an uncertain payment scheme (tournament), receiving feedback whether their performance was the highest or not (tournament) or not receiving such information (piece rate). Like in task 3, participants have to base their decisions on their beliefs about their relative performance in their group.

As such, this treatment will provide some insight into what extent choices between incentive schemes in task 3 are affected by factors that are also presented in task 4 compared to the unique factor that is missing in the task 4 choice, namely the desire or willingness to perform under a competitive incentive scheme.

Finally, participants provide information about their rank among the four players in each group in both the task 1 piece rate and the task 2 tournament performance (where a correct rank is rewarded by $1).

We find that women and men perform very similarly in both the piece rate scheme and the tournament scheme. The average performance in the piece rate is 10.15 for women and 10.68 for men, and in the tournament it is 11.8 and 12.1, respectively. There is no gender difference in performance. The increase in performance between the piece rate and the tournament seems to be due more to learning how to perform this task rather than increased effort in the tournament.[16] Of the 20 groups, 11 are won by women and 9 are won by men, and men and women with the same performance have the same probability of winning the tournament. Given the past tournament performance, 30% of women and 30% of men have higher earnings from a tournament scheme, which increases to 40% and 45%, respectively, when we add participants who are basically indifferent. However, in the experiment, 35% of women and 73% of men enter the tournament (a significant difference).

Figure 5.4a shows for each task 2 tournament performance quartile the proportion of participants who enter the tournament. Men have a higher chance to enter the tournament for any performance level.[17]

One driving factor could be that women and men differ in their beliefs about their relative performance in the tournament (explanation 2). In the experiment, 30 out of 40 men (75%!) believe that they were the best in their group of 4 (most of them were obviously wrong), that is men are highly overconfident. While women are also overconfident (17 (40%) believe they had the highest performance), men are significantly more overconfident than women. Can this gender difference account for differences in tournament entry?

Figure 5.4b shows the proportion of participants that enter the tournament as a function of their guessed rank in the task 2 tournament. Beliefs have a significant impact on the decision to enter the tournament, but gender differences remain even

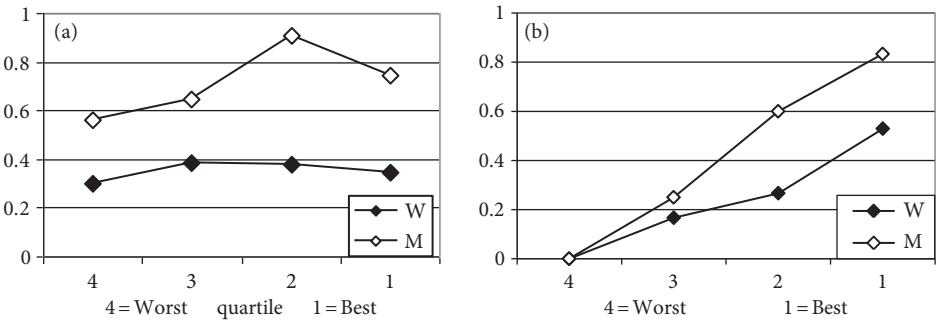


Figure 5.4. Proportion selecting tournament (Niederle and Vesterlund, 2007). (a) Conditional on initial tournament performance quartile. (b) Conditional on believed performance rank in initial tournament.

after controlling for beliefs, which account for about 30% of the initial gender gap in tournament entry (a result confirmed by regressions).

To study the impact of risk and feedback aversion on the decision to enter a tournament, we first study the decision in task 4 where participants decide whether to submit the task 1 piece rate performance to a piece rate or a tournament payment scheme. In this case the participants' actual performance, along with their beliefs about relative performance, can largely account for choices of women and men; the remaining gender gap in choices is economically small and not significant. That is gender differences do not follow the pattern found when choosing whether to enter a tournament and then perform. We studied a choice that mimics the decision of task 3, where participants decide whether to enter a tournament and then perform, only now, in task 4, participants did not have to perform anymore. Rather the payment was based on their past piece rate performance. When we eliminate the need for an upcoming tournament performance, the decisions of women and men can be entirely accounted for by their actual performance and their beliefs about their relative performance. This already casts doubt that risk and feedback aversion (explanations 3 and 4) are major factors in determining gender differences in choosing to enter a tournament. Furthermore, a regression on the task 3 decision of women and men to enter a tournament and then perform finds substantial gender differences, even when controlling for performance and beliefs and the choice in task 4.

In terms of money maximizing choices, high-performing women enter the tournament too little and low-performing men enter it too much (though by design their losses are smaller, as payments are dependent on performance). The result is that few women enter the competition and few women win the competition.

Experiments were useful in showing this gender difference, as they allowed for controls that would be hard to come by with labor data: Apart from being able to control performances in both environments, we could also ensure participants

that (a) there are no aspects of discrimination whatsoever, (b) their payments only depended on their decisions and performances of others, and (c) women were not treated differently than men. Furthermore, the experiments last less than 90 minutes; as such, any concerns about raising children are clearly not an issue. Meanwhile, many other experiments have replicated the basic result.[18]

Some final comments about the design choices in Niederle and Vesterlund (2007). For example, when trying to understand the impact of risk aversion, or aversion to receive information about whether one's performance was the best or not, we could have chosen different ways. For example, we could have tried to explicitly measure risk aversion, by having participants play various gambles, and measure their willingness to accept or reject those gambles. There are two comments to that approach.

The first concerns the actual risks or lotteries involved in the choices of participants. For example, for participants who have 14 or more correct answers, the chance of winning the tournament is 47% or higher. If participants maintain the performance after their choice of compensation scheme, the decision to enter the tournament becomes a gamble of receiving, per correct answer, either $2 with a probability of 47% (or more) or receiving 50 cents for sure. Hence, a gamble of a 47% chance of $28 (i.e., an expected value of $13) versus a sure gain of $7. Of all participants who solve 14 problems or more, 8/12 of the women and 3/12 of the men do not take this gamble.[19] Similarly, for participants who have 11 or fewer correct answers, the chance of winning the tournament is 5.6% or less. Thus entering the tournament means receiving $2 per correct answer with a probability of 5.6% (or less) versus receiving 50 cents for sure. For all participants who solve 11 correct answers, this is a choice between a 5.6% chance of winning $22 (i.e., an expected value of $1.23) compared to receiving $5.5 for sure. Of the men who solve 11 problems or less, 11 of 18 take this gamble while only 5 of 17 women do.[20] I am not aware of any studies that find such extreme gender differences in risk aversion. Furthermore, if risk aversion is the most important explanation for the gender gap in tournament entry, men should not enter the tournament with a higher probability than women for all performance levels, but rather the entry decision of women should be shifted to the right of that of men.

Second, even if we found that more risk-averse participants enter the tournaments at a lower propensity, it is not clear that risk aversion may indeed be the explaining factor. Similarly, if we found that risk aversion on its own does not reduce the gender gap, it could be that we measured risk attitudes on the wrong lotteries, since perceived lotteries may be very different.

Basically, the problem is that choices of participants depend in a very intricate way on risk aversion, which may be hard to capture, As such, indirect approaches like the ones we took in our chapter, may be more reliable, as they circumvent the issue of measurement, or making the right assumptions about how risk aversion enters the decision precisely (in combination with beliefs about performance, actual performance, etc.).

Conclusions

In this chapter, I wanted to show how to use experiments to test theory. I showed how the theory can be tested at deeper levels, by attacking the assumptions of the theory directly. I also provided two examples of design to test the theory: the two-way design and the elimination design. In the two-way design, the initial environment, which allowed for two competing theories to explain the initial results, is changed in a way to separate the two theories. In the elimination design, the explanatory power of a theory for a phenomenon is questioned by changing the environment in a way that (a) the theory has no explanatory power anymore and (b) the phenomenon is still present. This at least casts doubt that the initial theory was the main driving factor for the result.

Finally, intelligent design can also be used when the initial hypothesis is not grounded in a careful model. In fact, most examples of intelligent design I presented in this chapter do not really deeply rely on the parameters of a model, but rather exploit broad results or assumptions. Often a more direct approach may help us to learn more.

NOTES

1. For example, Econometrica states that "We understand that there may be a need for exceptions to the policy for confidentiality or other reasons. When this is the case, it must be clearly stated at the time of submission that certain data or other appendix materials cannot or will not be made available, and an explanation provided. Such exceptions require prior approval by the editors."
http://www.econometricsociety.org/submissions.asp#Experimental, accessed 8/11/2010. For an overview on some of the arguments, see Hertwig and Ortmann (2008).

2. On a similar note, cognitive hierarchies (see Camerer et al. (2004)) assume that a level 0 player plays a random strategy, and a level k player best responds to a distribution of players of levels $k - 1$ and lower, where the distribution is given by the restriction of a common Poisson distribution over all levels on the levels strictly below k.

3. See Bosch-Domènech et al. (2002) for a beauty contest game played in the field, and also see Ho et al. (1998), Costa-Gomes and Crawford (2006), Crawford and Iriberri (2007), Costa-Gomes et al. (2009), and also Stahl (1998).

4. Some papers, such as Costa-Gomes and Crawford (2006), use a somewhat more stringent test by taking into account not only the actions of players, but also what information they used when making a decision. Specifically, parameters of the game were hidden and had to be actively uncovered by participants. The pattern of lookups can exclude misspecification of certain "random" strategies when the data needed to be a k-level player was not even uncovered. Camerer et al. (2004) propose to fit data using a unique parameter that determines the distribution of types.

5. However, there has not yet been too much of a debate about the predictive power of k-level models, and what behavior is easily excluded, even though such debates are

active for other models of deviations of rational behavior, such as Quantal Response Equilibrium (see McKelvey and Palfrey (1995) and Haile Hortaçsu and Kosenok (2008)).

6. However, study applies to any belief-based explanation of the winners' curse. This includes, for example, cursed equilibrium (see Eyster and Rabin (2005)), and analogy-based expectation equilibrium (Jehiel, 2005; Jehiel and Koessler, 2008).

7. For the table, we use the following, slightly different classification: Underbid/Signal-bid/Overbid/Above-10-bid are bid of, respectively, (i) $b < x - 0.25$, (ii) $x - 0.25 \leq b \leq x + 0.25$, (iii) $x + 0.25 < b \leq 10$, and (iv) $b > 10$, with the exception for $x = 10$, where a bid b of $9.75 \leq b \leq 10.25$ falls in category (ii), and only a bid $b > 10.25$ falls in category (iv).

8. In their common-value auctions, the value of the item x is a random uniform draw from [\$50, \$950], where each of the six bidders receives a private signal y, drawn independently from $[x - \$15, x + \$15]$. Because of boundaries, when attention is restricted to x in [65, 935], the bid factor (the amount to be deducted from the signal) of the RNNE is about 15, which is close to the loss-free strategy (where bidders can ensure never to lose money). The bid factor for the break-even strategy (the strategy that yields zero expected profits with occasional looses) is about 10.71.

9. See Friedman (1992), Kagel and Roth (1992), Cox et al. (1992), Merlo and Schotter (1992), and Harrison (1992).

10. See Roth (1994) about what constitutes an experiment, and his argument to keep up the integrity about how much "search" went on in terms of how many (unreported) trials or treatments were run in order to find a specific result.

11. Eckel and Grossman (2002) and Croson and Gneezy (2009) summarize the experimental literature in economics and conclude that women exhibit greater risk aversion in choices. A summary of the psychology literature is presented by Byrnes et al. (1999). They provide a meta-analysis of 150 risk experiments and demonstrate that while women in some situations are significantly more averse to risk, many studies find no gender difference.

12. For example, Mobius et al. (2010) explicitly ask participants about their willingness to pay (or get compensated for) receiving information about their performance in an IQ-like quiz. We find that men are significantly less averse to receiving feedback than women.

13. We did not want to trigger any demand effects or psychological biases such as priming by pointing out that we study gender.

14. On a technical note, by paying the tournament winner by performance rather than a fixed prize, we avoid providing information about winning performances, or distorting incentives for very high-performing individuals.

15. There is a large literature on the debate whether women are more altruistic than men and hence may be more or less worried about imposing a negative externality on other participants (see Croson and Gneezy (2009) and Andreoni and Vesterlund (2001)).

16. This is supported by the fact that changes in performance between task 2 and task 3 are independent of the chosen incentive scheme in task 3. Note that this does not imply that participants do never provide effort, rather it appears their baseline effort is already quite high.

17. Similar results are obtained when we consider the performance after the entry decision, rather than the one before the entry decision.

18. For an overview see, for example, Niederle and Vesterlund (2010).

19. This difference is marginally significant with a two-sided Fisher's exact test ($p = 0.100$).

20. This difference is marginally significant with a two-sided Fisher's exact test ($p = 0.092$).

REFERENCES

Andreoni, J. and L. Vesterlund. 2001. Which is the Fair Sex? On Gender Differences in Altruism. *Quarterly Journal of Economics* **116**:293–312.

Beyer, S. 1990. Gender Differences in the Accuracy of Self-Evaluations of Performance. *Journal of Personality and Social Psychology* **LIX**:960–970.

Beyer, S. and E. M. Bowden. 1997. Gender Differences in Self-Perceptions: Convergent Evidence from Three Measures of Accuracy and Bias. *Personality and Social Psychology Bulletin* **XXIII**:157–172.

Bosch-Domènech, A., J. García-Montalvo, R. Nagel, and A. Satorra. 2002. One, Two, (Three), Infinity…: Newspaper and Lab Beauty-Contest Experiments. *American Economic Review* **92**(5):1687–1701.

Byrnes, J. P., D. C. Miller, and W. D. Schafer. 1990. Gender Differences in Risk Taking: A Meta-Analysis. *Psychological Bulletin* **LXXV**:367–383.

Camerer, C., T. Ho, and J. Chong. 2004. A Cognitive Hierarchy Model of One-Shot Games. *Quarterly Journal of Economics* **119**(3):861–898.

Camerer, C. and D. Lovallo. 1999. Overconfidence and Excess Entry: An Experimental Approach. *American Economic Review* **89**(1):306–318.

Casari, M., J. C. Ham, and J. H. Kagel. 2007. Selection Bias, Demographic Effects, and Ability Effects in Common Value Auction Experiments. *American Economic Review* **97**(4):1278–1304.

Costa-Gomes, M. A. and V. P. Crawford. 2006. Cognition and Behavior in Two-Person Guessing Games: An Experimental Study. *American Economic Review* **96**:1737–1768.

Costa-Gomes, M. A., V. P. Crawford, and N. Iriberri. 2009. Comparing Models of Strategic Thinking in Van Huyck, Battalio, and Beil's Coordination Games. *Journal of the European Economic Association* **7**:377–387.

Costa-Gomes, M. A. and G. Weizsaecker. 2008. Stated Beliefs and Play in Normal Form Games. *Review of Economic Studies* **75**:729–762.

Cox, J. C., V. L. Smith, and J. M. Walker. 1988. Theory and Individual Behavior of First-Price Auctions. *Journal of Risk and Uncertainty* **1**:61–99.

Cox, J. C., V. L. Smith, and J. M. Walker. 1992. Theory and Misbehavior of First-Price Auctions: Comment. *American Economic Review* **82**(5):1392–1412.

Crawford, V. P. and N. Iriberri. 2007. Level-*k* Auctions: Can a Non-Equilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions. *Econometrica* **75**:1721–1770.

Croson, R. and U. Gneezy. 2009. Gender Differences in Preferences. *Journal of Economic Literature* **47**(2):448–474.

Eckel, C. C. and P. J. Grossman. 2002. Sex and Risk: Experimental Evidence. In *Handbook of Experimental Economics Results*, eds. C. Plott and V. Smith. Amsterdam: Elsevier Science B.V./North-Holland.

Erev, I., E. Ert, and A. E. Roth. 2010a. A Choice Prediction Competition for Market Entry Games: An Introduction. *Games* **1**(2):117–136.

Erev, I., E. Ert, A. E. Roth, E. Haruvy, S. Herzog, R. Hau, R. Hertwig, T. Steward, R. West, and C. Lebiere. 2010b. A Choice Prediction Competition, for Choices from Experience and from Description. *Journal of Behavioral Decision Making* **23**:15–47.

Erev, I. and A.E. Roth. 1998. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review* **88**(4):848–881.

Erev, I., A. E. Roth, R. L. Slonim, and G. Barron. 2007. Learning and Equilibrium as Useful Approximations: Accuracy of Prediction on Randomly Selected Constant Sum Games. *Economic Theory* **33**:29–51.

Eyster, E. and M. Rabin. 2005. Cursed Equilibrium. *Econometrica* **73**(5):1623–1672.

Friedman, D. 1992. Theory and Misbehavior of First-Price Auctions: Comment. *American Economic Review* **82**(5):1374–1378.

Georganas, S., P. J. Healy and R. Weber. 2009. On the Persistence of Strategic Sophistication. Unpublished.

Gneezy, U., M. Niederle, and A. Rustichini. 2003. Performance in Competitive Environments: Gender Differences. *Quarterly Journal of Economics* **CXVIII**:1049–1074.

Grosskopf, B. and R. Nagel. 2008. The Two-Person Beauty Contest. *Games and Economic Behavior* **62**:93–99.

Haile, P., A. Hortaçsu, and G. Kosenok. 2008. On the Empirical Content of Quantal Response Equilibrium. *American Economic Review* **98**(1):180–200.

Harrison, G. W. 1989. Theory and Misbehavior of First-Price Auctions. *American Economic Review* **79**(4):749–762.

Harrison, G. W. 1992. Theory and Misbehavior of First-Price Auctions: Reply. *American Economic Review* **82**(5):1426–1443.

Hertwig, R. and A. Ortmann. 2001. Experimental Practices in Economics: A Methodological Challenge for Psychologists? *Behavioral and Brain Sciences* **24**:383–451.

Ho, T., C. Camerer, and K. Weigelt. 1998. Iterated Dominance and Iterated Best response in *p*-Beauty Contests. *American Economic Review* **LXXXVIII**:947–969.

Hoelzl, E. and A. Rustichini. 2005. Overconfident: Do You Put Your Money on It? *Economic Journal* **115**(503):305–318.

Ivanov, A., D. Levin, and M. Niederle. 2010. Can Relaxation of Beliefs Rationalize the Winner's Curse? An Experimental Study. *Econometrica* **78**(4):1435–1452.

Jehiel, P. 2005. Analogy-Based Expectation Equilibrium. *Journal of Economic Theory* **123**:81–104.

Jehiel, P. and F. Koessler. 2008. Revisiting Games of Incomplete Information with Analogy-Based Expectations. *Games and Economic Behavior* **62**:533–557.

Kagel, J. H. 1995. Auction: Survey of Experimental Research. In *The Handbook of Experimental Economics*, eds. A. E. Roth and J. H. Kagel. Princeton, NJ: Princeton University Press.

Kagel, J. H. and D. Levin. 1993. Independent Private Value Auctions: Bidder Behavior in First-, Second-, and Third-Price Auctions with Varying Numbers of Bidders. *The Economic Journal* **103**(419):868–879.

Kagel, J. H. and D. Levin. Forthcoming. Auctions: A Survey of Experimental Research, 1995-2008. In *The Handbook of Experimental Economics*, Volume 2, eds. A. E. Roth and J. H. Kagel. Princeton, NJ: Princeton University Press.

Kagel, J. H. and A. E. Roth. 1992. Theory and Misbehavior in First-Price Auctions: Comment. *American Economic Review* **82**(5):1379–1391.
Lichtenstein, S., B. Fischhoff, and L. Phillips. 1982. Calibration and Probabilities: The State of the Art to 1980. In *Judgment under Uncertainty: Heurstics and Biases*, eds. D. Kahneman, P. Slovic, and A. Tversky. New York: Cambridge University Press.
McKelvey, R. D. and T. R. Palfrey. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* **10**(1):6–38.
Merlo, A. and A. Schotter. 1992. Theory and Misbehavior of First-Price Auctions: Comment. *American Economic Review* **82**:1413–1425.
Mobius, M., M. Niederle, P. Niehaus, and T. Rosenblat. 2010. Maintaining Self-Confidence: Theory and Experimental Evidence. Unpublished.
Nagel, R. 1995. Unraveling in Guessing Games: An Experimental Study. *American Economic Review* **85**(5):1313–1326.
Niederle, M. and A. E. Roth. 2009. Market Culture: How Rules Governing Exploding Offers Affect Market Performance. *American Economic Journal: Microeconomics* **1**(2):199–219.
Niederle, M. and L. Vesterlund. 2007. Do Women Shy Away from Competition? Do Men Compete too Much? *Quarterly Journal of Economics* **122**(3):1067–1101.
Niederle, M. and L. Vesterlund. 2010. Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives* **24**(2):129–144.
Roth, A. E. 1994. Let's Keep the Con Out of Experimental Econ.: A Methodological Note. *Empirical Economics* **19**:279–289.
Roth, A. E. 2008. *What Have We Learned from Market Design*? Hahn Lecture, *Economic Journal* **118**:285–310.
Rubinstein, A. 2006. Dilemmas of an Economic Theorist. *Econometrica* **74**:865–883.
Stahl, D. 1998. Is Step-*j* Thinking an Arbitrary Modeling Restriction or a Fact of Human Nature? *Journal of Economic Behavior and Organization* **XXXVII**:33–51.
Stahl, D. and P. Wilson. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior* **10**:218–254.

CHAPTER 6

THE INTERPLAY BETWEEN THEORY AND EXPERIMENTS

LEEAT YARIV

Introduction

When experimentalists who are theory-oriented or theorists who frequently run experiments discuss the interplay between theory and experiments, there is great danger for an optimism bias. Such scholars are often friendly to both theory and experiments and hold a desire to see theory and experiments flourish hand in hand. They also tend to self-select into conferences and seminar series that are friendly to both these fields. Admitting to this potential bias, my goal is to inspect some metadata on the recent evolution of the fields of theory and experiments in order to get an empirically sound impression.

I report observations regarding publications in the top general interest and field journals during 1998, 2003, and 2008. I also use data on the faculty coauthoring these papers, documenting the rank of the schools they are affiliated with (top 15 or below top 15) and their fields of interest (pure theory, pure experiments, theory and experiments).

There are two realms of results that the data point to. With respect to papers, there do not seem to be significant increases in the volume of experimental work that is being published, be it with or without a theoretical spin. Field journals, however, do exhibit some significant time trends, with more experimental publications over time.

With respect to authors, I find that theorists and experimentalists publishing at the leading journals have rather similar profiles, with two exceptions. First, well-published theorists are slightly better placed than their experimental counterparts. Second, when looking at the coauthorship networks, experimentalists are slightly more connected than theorists. In addition, there is a substantial time trend in terms of theorists who dabble in experimental work; they are significantly increasing in volume over the time period the data is from, particularly for faculty in top-ranked institutions.

The Data

Data were collected on microeconomics publications in 1998, 2003, and 2008 in the top general interest economics journals: *American Economic Review* (*AER*), *Econometrica* (*EMA*), *Journal of Political Economy* (*JPE*), *Quarterly Journal of Economics* (*QJE*), and *Review of Economic Studies* (*ReSTUD*), as well as the leading field journals in microeconomics (in existence since 1998): *Games and Economic Behavior* (*GEB*) and *Journal of Economic Theory* (*JET*).[1] Each paper was classified into one of three categories: theory, experiments, or theory and experiments.

In addition, data were collected on the publication records, specific fields of interest, and school rankings of their microeconomics faculty as of 2008. Publication records and fields of interest were harvested from faculty webpages. School rankings were categorized coarsely into top 15 and below top 15.[2]

Publication Trends

The Papers

The first place to look for general time trends is the dynamics of the publication process. Table 6.1 contains the split between papers classified as theoretical (absent

Table 6.1. Publication Time Trends

	Year	Overall	Theory	Experiments	Theory-Based Experiments
Top Journals	1998	242	60	10	8
	2003	279	62	7	1
	2008	240	61	11	6
Field Journals	1998	140	133	7 (1 JET, 6 GEB)	6
	2003	171	159	12 (2 JET, 10 GEB)	7
	2008	160	144	16 (all GEB)	14

any use of newly collected laboratory data), experimental (absent any suggested theoretical model), and experimental based on some theoretical modeling.[3]

The top panel of Table 6.1 suggests very limited time trends, in terms of both absolute and relative volumes of published experimental work. The fraction of experimental papers accounts for less than 5% of all published papers in the top general interest journals, as well as around 15% of the published microeconomics papers (using either theoretical or experimental methodologies). In fact, most experimental papers are published in *EMA* and *AER*: 8 of 10 in 1998, 4 of 7 in 2003, and 8 of 11 in 2008.[4]

The trends pertaining to the top field journals are reported in the bottom panel of Table 6.1 and exhibit somewhat different features. In both absolute and relative terms, the number of experimental papers published in these two journals has increased, particularly in *GEB*. Of these, both the number and the fraction of theory-based experiments have risen.

There are different interpretations one can give to these observations. It might be that a new field first gains acceptance in the top journals, which may have occurred for the experimental literature by the mid-1990s, and then *trickles down* to the more specialized journals. Alternatively, it might be that a field gets established through publications in specialized journals first, so that what we are observing is a *trickling up* through the journal rankings of experimental work. Of course, it may also be that the three years reported in this chapter are not fully representative. Further data collection as well as historical analyses of other fields' emergence would be useful in assessing the likelihood of each of these theories.

The Authors

Another dimension by which to inspect the time trends of the publication process is through the affiliation of the authors. It is conceivable that papers in emerging fields require authors to have some stamps of quality (say, an affiliation to a highly ranked school) in order to get published. Figure 6.1 illustrates two ways by which to inspect these effects.

In the top panel of Figure 6.1, the percentage of authors in the top 15 schools publishing in the top general interest and field journals are reported for the three time periods inspected. The bottom panel considers the percentage of papers with at least one author in a top 15 school in the two classes of journals considered (there are too few data points to be reported for the set of theory-based experimental papers).

The figure suggests several trends. First, the percentage of authors coming from the very top institutions and writing theory-oriented papers has risen within the general interest journals, but has decreased for experimental publications and for all types of papers in the field journals. Second, this image is reversed when looking at the need to have at least one author from one of the very top institutions in the general interest journals. In these journals, experimental papers appear to

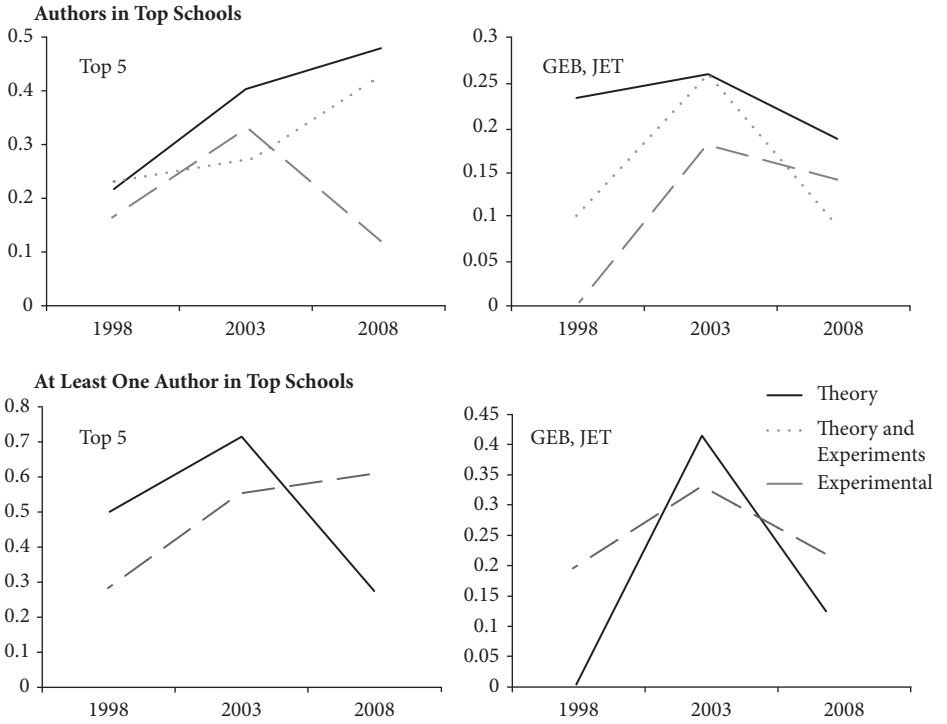


Figure 6.1. Author time trends.

increasingly be associated with at least one author from a top 15 school, while for theory papers the opposite trend is observed. For the field journals, the general trend remains, though the decreases are more pronounced when looking at the need for one author in a top school.

These observations are potentially driven by forces pertaining to both the demand and the supply of papers with different characteristics. On the demand side, journals may have greater tendencies to accept papers with particular profiles of coauthors (that are expected to be associated with papers' quality). For instance, if general interest journals had greater propensity to accept papers with at least one coauthor affiliated with a top 15 school, much of the trends that are observed would be explained.

On the supply side, there is evidence suggesting that the structure of coauthorships has changed over time (see, e.g., Gans (2001), Ellison (2002), and Einav and Yariv (2006), as well as the discussion in the following sections). In particular, there is an acknowledged trend in Economics for fundamental papers to involve an increasing number of coauthors. Furthermore, it is conceivable that theorists coauthor with faculty in their own school more frequently than experimentalists, simply due to the fact that most of the leading universities do not employ an abundance of experimentalists. These two forces would generate an increasing wedge

between the fraction of authors in the very top institutions that collaborate on leading theoretical and experimental work.

In what follows, I discuss some of the attributes corresponding to the coauthorship structures of theorists and experimentalists, the supply side of the market in question.

Coauthorship Trends

Over the time period covered in this chapter, looking at authors of papers in the top journals, the percentage of authors who were pure experimentalists increased over time (from 5% in 1998, to 7% in 2003, to 11% in 2008), while the percentage of authors who were pure theorists remained stable (at 69% in 1998 and 70% in 2003 and 2008), thereby implying that the complementary fraction of authors who combined experiments and theory in their work declined.

Figure 6.2 describes the network of coauthorships within the top general interest journals in the three years that were inspected.[5] Black nodes correspond to theorists, white nodes to experimentalists, and gray nodes to researchers who engage in a substantial number of both theoretical and experimental projects. A link between two nodes implies that the respective faculty have appeared as coauthors in a published paper. Solid links represent coauthorship on a theoretical paper,

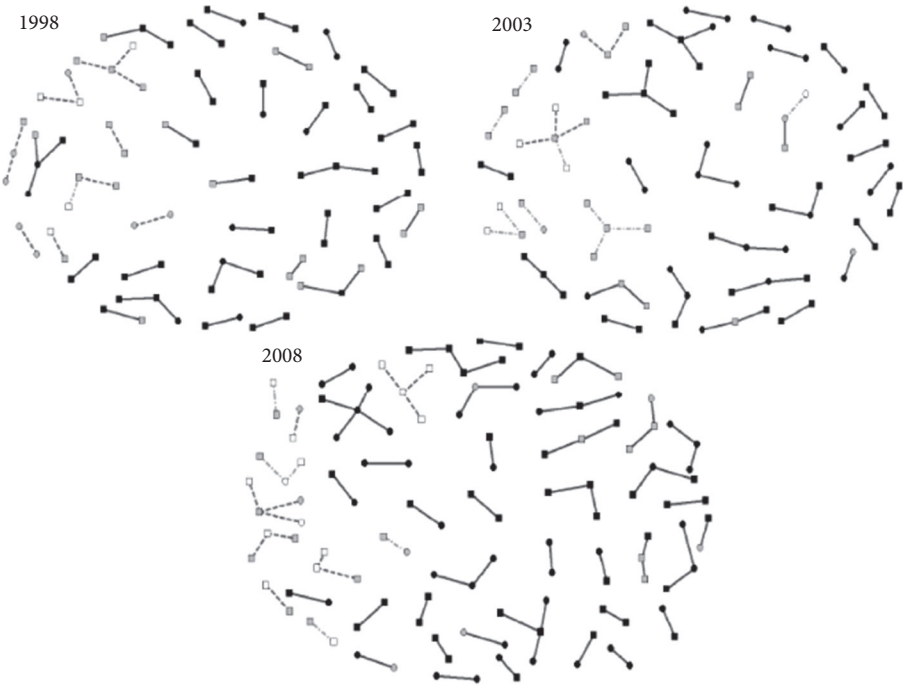


Figure 6.2. Coauthorships across time.

dashed links represent coauthorship on an experimental paper, and links comprised of dashes and dots correspond to coauthorships on papers that fall within the theory-based experiments category.[6]

Looking at Figure 6.2, one may notice the increase of the volume of nodes over the years, as well as the rise in the number of connected components, mostly star-shaped, involving three or more authors.

The increase in volume of nodes is presumably due in big part to the increase in the number of coauthors involved in each paper. Indeed, the fraction of 1998 papers discussed here that were solo-authored was 40%, and the fraction involving two coauthors was 47%. In 2003 these numbers were 38% and 45%, respectively; and in 2008 they were 26% and 47%. This general trend has been documented with more extensive data sets in Gans (2001), Ellison (2002), and Einav and Yariv (2006).

We now turn to the differences between theorists and experimentalists. Figure 6.3 depicts the cumulative distributions of degrees (number of connections in the network) of theorists and experimentalists across time (where, for the sake of the figure, experimentalists are taken as those engaged in a substantial amount of experimental work, corresponding to either the white or gray nodes in Figure 6.2).

While the differences between the cumulative distributions corresponding to theorists and experimentalists are small, the distribution corresponding to experimentalists consistently first-order stochastically dominates that corresponding to theorists. Furthermore, there are not many significant time trends (though the number of solo-authored papers has decreased significantly, which is mirrored by the fraction of degree 1 agents plummeting over time).[7]

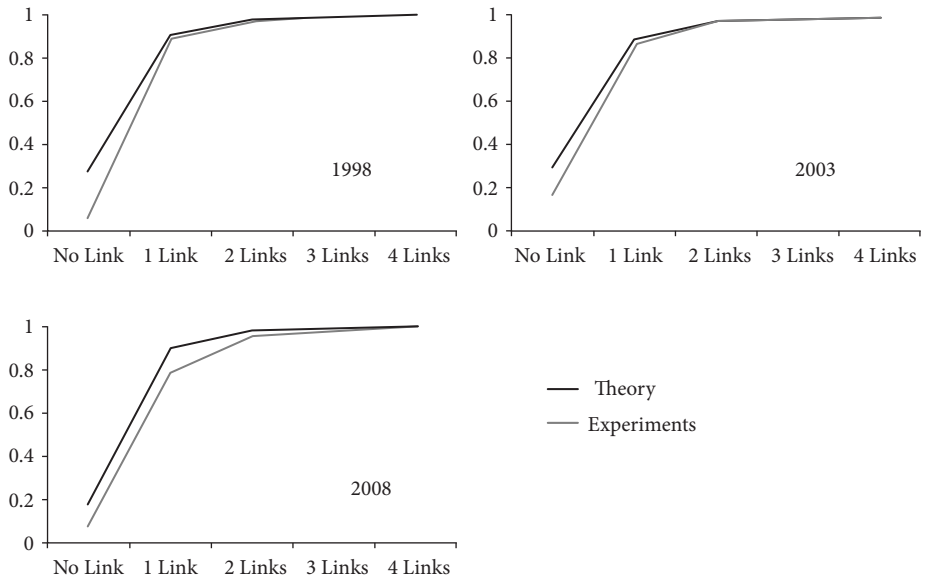


Figure 6.3. Cumulative degree distributions for theorists and experimentalists.

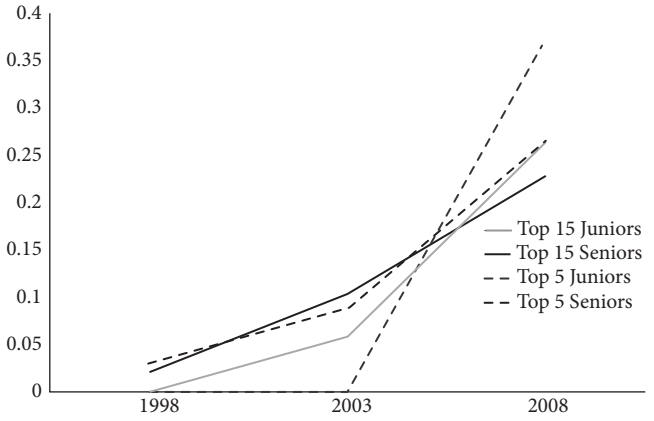


Figure 6.4. Dabbling theorists across time.

Theorists Dabbling in Experiments

I now turn to look at the emergence of theorists who are seriously interested in experimental work enough that they have been involved in one or two experimental projects themselves. Focusing on current affiliations, Figure 6.4 documents the percentage of theorists having dabbled in experiments by each of the years inspected in this chapter.[8] The figure depicts the trends for the top 15 schools and top 5 schools, for senior and junior faculty.[9]

Figure 6.4 illustrates a substantial increase in the number of theorists dabbling in experiments, particularly among junior faculty in recent years. This trend is noticeably more pronounced for the higher ranked schools.

It is often believed that junior faculty can take less risks in the choice of topics they work on, having impending tenure decisions in sight. In that respect, the comparison between junior and senior theorists is interesting. In 1998 and 2003, junior theorists were less likely to engage in one-time experiments than senior theorists. Nonetheless, in 2008, this observation is reversed. To the extent that juniors' field choices are suggestive of how risky these fields are perceived to be, experiments may be thought of as more mainstream by theorists nowadays.

Conclusion

In order to assess the interplay between theory and experiments, I used data on publications and microeconomics faculty in three years representing the last decade: 1998, 2003, and 2008.

There are several important messages that one can take from this analysis. Regarding publications, the percentage of experimental papers has not changed significantly at the top general interest journals, but increased substantially within

the top field journals. Regarding the researchers' profiles, experimentalists and theorists have similar characteristics. Experimentalists do seem slightly more connected, and well-published theorists (namely, those publishing in the top journals) are slightly better placed in terms of the ranking of the school they are affiliated with. Theorists and experimentalists do interact. In fact, there is a dramatic increase in the recent volume of theorists dabbling in experiments.

These observations are impressionistic in that they are based on limited observations from only three years. It will be interesting to continue following these trends as years go by.

NOTES

I thank Muruvvet Buyukboyaci for excellent research assistance. Financial support from the National Science Foundation (SES 0551014) and the Gordon and Betty Moore Foundation is gratefully acknowledged.

1. Several recent journals focusing on microeconomics work have emerged over the past decade that are clearly perceived as leading outlets. For the sake of the dynamic analysis performed in this chapter, only journals that were in existence since 1998 were considered.

2. The ranking of the top 15 economics units worldwide was taken as (1) Harvard, (2) Massachusetts Institute of Technology, (3) Stanford, (4) Princeton, (5) University of Chicago, (6) University of Pennsylvania, (7) Northwestern, (8) New York University, (9) Yale, (10) London School of Economics, (11) University of California, Los Angeles, (12) University of California, Berkeley, (13) University of Minnesota, (14) University of Rochester, (15) University of Michigan. Faculty affiliated with any of these universities (from any department or professional school) were classified as belonging to a "top 15" school. This ranking is taken as an interpolation of an array of available rankings (see Amir and Knauff (2005) and Einav and Yariv (2006)).

3. A paper was classified into the last category (theory-based experiments) whenever there was an explicit mention of theoretical analysis.

4. The reported figures pertain to full papers only. In fact, over the three inspected years, only one experimental piece was published as a note in *AER*.

5. Similar qualitative insights to those described in this section emerge when considering field journals as well.

6. In our data, no two faculty were involved in two papers belonging to different categories and so the classification of links is well-defined.

7. As mentioned before, the observations reported here remain qualitatively similar when looking at the field journals, with one exception. When looking at degree distributions, in 1998, that corresponding to experimentalists was stochastically dominated by that corresponding to theorists within *JET* and *GEB*.

8. This should be interpreted with care, as some of these faculty may have changed affiliations over the years. Furthermore, dabbling in experiments often turns into a career pursuit, and while those theorists who turned experimentalists in the late 1990s are not

counted as "dabblers," those who will become experimentalists after their experimentation in 2008 are counted as such.

9. Following the ranking used throughout the chapter, the top five schools are taken to be: (1) Harvard, (2) Massachusetts Institute of Technology, (3) Stanford, (4) Princeton, and (5) University of Chicago.

REFERENCES

Amir, R. and M. Knauff. 2005. Ranking Economics Departments Worldwide on the Basis of PhD Placement. CORE Discussion Paper 2005/51.

Einav, L. and L. Yariv. 2006. What's in a Surname? The Effects of Surname Initials on Academic Success. *The Journal of Economic Perspectives* **20**(1):175–188.

Ellison, G. 2002. The Slowdown of the Economics Publishing Process. *Journal of Political Economy* **105**(5):947–993.

Gans, J. S., ed. 2001. *Publishing Economics: Analyses of the Academic Journal Market in Economics*. Cheltenham, UK: Edward Elgar Publishing.

CHAPTER 7

MAXIMS FOR EXPERIMENTERS

MARTIN DUFWENBERG

Introduction

How do you produce a good experiment? How should you write up the results? The answers may depend on the goal of the exercise. While there may be more than one sensible formula, I have my clear favorite. I present it here, alongside four corollaries on what not to do if one follows the formula. I also make a related proposal for how to submit papers to journals, with the results in a sealed envelope. I suspect that my call for such submissions will not be heeded, but I have a backup plan and therefore finally make a request to editors and referees who evaluate experimental work.

The Formula

Economists are interested in how the world works. They wish to understand, for example, the causes and effects of unemployment, how the rules of the Nasdaq effects stock prices, and how vengeance may influence holdup in business partnerships. Since economic outcomes are shaped by how humans reason as well as the principles by which groups of individuals reach outcomes in games, economists care about these matters too.

Theorists tell stories about how the world works. It is natural to wonder what stories are empirically relevant. In this connection, experiments may be useful,

as they can help provide answers. My advice to students who want to do such experimental work is this: Follow the formula

$$I \rightarrow D \rightarrow H \rightarrow R$$

where:

- I stands for *idea*—a statement about how the world works, often but not necessarily derived from theory.
- D stands for *design*—a set of experimental conditions and treatments meant to be optimal for shedding light on the empirical relevance of the idea.
- H stands for *hypothesis*—a statement formulated in terms of the data generated by the design, which would support the empirical relevance of the idea.
- Finally, R stands for *result*—the evaluation of whether or not the data generated by the design support the hypothesis and the idea.

What Not to Do

The formula comes with four corollaries regarding what not to do:

- Don't get (too) curious about changes to the design!

The point here is not to confuse interest in the design with interest in the idea. Under the formula, the design has its raison d'être only as a tool for shedding light on the empirical relevance of the idea. Therefore, making a change to the original design to see what happens is not useful, unless one somehow can argue that this change would shed light on the idea.

- Don't think it is always desirable to "explain" the data!

Under the formula, the goal of your endeavor is to test the idea, not to understand the data. Suppose, for example, that the idea you test is a theory which abstracts away from some aspects (or 'frictions') known to be relevant to some extent. Suppose the data provide some but not perfect support for the theory. You may be done, if explaining the frictions were never part of your research goal. Or suppose the data do not support the theory. That may be an interesting finding, and again you may be done. I'm not saying that explaining the data is never useful, only that one shouldn't take for granted that it is.

- Don't decide on the idea after seeing the data!

Since the design is a response to the desire to test an idea, it is meaningless to run a design without an idea to test (if one follows the formula). And if ideas are specified ex post, there are two problems: First, it is unlikely that the design will be

optimal for testing the idea, since the design was not constructed with the idea in mind. Second, there is risk of outcome bias in the perception of causality.

- Don't worry about the data!

Running an experiment is similar to decision making under uncertainty. One wants to make decisions that maximize expected utility, and in an uncertain world one can't rationally always hope to make the decisions that turns out to be best ex post. For example, drawing to an inside straight at poker without proper odds is a sucker play that loses money in the long run, even if every now and then the straight is made. Similarly, experimenters should run the experiment they deem to have the greatest scientific merit viewed from an ex ante perspective. With respect to testing theory, the value attached to the research you are doing (in terms of publishability) shouldn't depend on the nature of the data.

This last corollary comes with a caveat. What if the editors and readership of our journals are more interested in certain kinds of results than others, say results that are deemed surprising? This may skew researchers' incentives, as maximizing the probability of a surprising result need not be the same thing as choosing the project with the highest expected scientific value from an ex ante point of view. And if nonsurprising results do not make it into journals, this may skew the picture of how the world works that emerges from published research. These problems lead me to a proposal:

The Sealed-Envelope-Submission Strategy

When you write up a research paper and submit it to a journal for publication . . .

- Don't mention results in the abstract!
- Don't mention results in the introduction!
- In fact, don't mention the results at all in the submitted paper. Put the results in a sealed envelope! Ask the editor and the referees to make their call before opening the envelope. Only once they have decided whether or not to publish the paper, they may open the envelope, study the data, and read your summary.

A Request to Editors and Referees

Will journals give serious consideration to submissions with the results in sealed envelopes? Will editors resist the temptation to sneak a peak? Might they even be convinced to insist that manuscripts be submitted this way? I believe the answer

is: probably not. . . . If so, then my call to editors and referees is this: *Evaluate the research you consider as if you had to form an opinion before looking at the data!*

NOTES

I thank Gary Charness for helpful comments. After this note was completed, Larry Samuelson pointed out to me that proposals similar to my sealed-envelope-submission strategy have been made in earlier work that discusses the hazards of a publication bias in favor of statistically significant results; see, for example, Rosenthal (1966), Walster and Cleary (1971), and Feige (1975).

REFERENCES

Feige, E. 1975. The Consequences of Journal Editorial Policies and a Suggestion for Revision. *The Journal of Political Economy* **83**:1291–1296.

Rosenthal, R. 1966. *Experimenter Effects in Behavioral Research*. New York: Appleton-Century-Croft.

Walster, G. and T. Cleary. 1970. A Proposal for a New Editorial Policy in the Social Sciences. *The American Statistician* **24**:16–19.

CHAPTER 8

WHAT IS AN ECONOMIC THEORY THAT CAN INFORM EXPERIMENTS?

URI GNEEZY AND PEDRO REY-BIEL

Economic models abstract from complex human behavior in a way that sheds some insight into a particular aspect of such behavior. This process inherently ignores important aspects of the real world. Similarly, experimental design frequently uses abstraction in order to be able to do comparative statics and reduce the number of possible explanations. Therefore, an important role of both theory and experiments is to shed light on behavior using simplified versions of the world.

Theories and experiments could and should, in many cases, inform each other. An economic theory is more useful if it not only is an intellectual exercise but also relates to empirical relevant behavior. Experiments that are based on a set of alternative, well-defined hypotheses are more useful.

Yet, testing theories is not the only role of experiments. First, it is not clear what constitutes an economic theory. In the current state of the profession, an economic theory is a precise, nontrivial, mathematical model. This is not the case in other social sciences, where theory could come in other forms such as a verbal description or a flaw chart. We argue that although mathematical models have dominated the neo-classical economic discussion since the 1950s, we should not restrict ourselves to this form.

Indeed, this was not the case in the past. The founding fathers of economics as an academic profession (e.g., Adam Smith, David Ricardo, François Quesnay) were considered philosophers at the time and used verbal arguments. The concepts of the invisible hand or the theory of the relative comparative advantage are still enormously useful in explaining why competitive markets and specialization work and how trade is generated.

Mathematical formalism has been successful in economics because it unifies scientific approaches, avoiding having different explanations for each possible manifestation of similar economic phenomena. In that sense, Economics has been partially able to avoid the problem of other social sciences, in which often theories are used like a toothbrushes: Everyone has one . . . and do not like to use anybody else's. Our argument is that economists should take advantage of this unifying approach in order to produce stronger theories, but they should not constrain themselves to the limited toolbox of using mathematical models exclusively.

Take, for example, the recent literature on gender and competition (Gneezy et al. 2003; Niederle 2015, Chapter 5 this volume). The experimental papers do not contain a single mathematical formula. Yet, they are inspired by theory—Darwin's (1871) sexual selection theory, showing that certain evolutionary traits, both physical and with respect to attitudes, can be explained by competition. Subsequent similar theories by Bateman (1948) and Trivers (1972) state that differences in competitiveness may have evolved because of competition for sex, where promiscuity is more valuable to the reproductive success of males than females.

This example shows that economic experiments do not necessarily need to ground their theory in economics and that theory from other areas, such as biology, psychology, or anthropology, can be important in explaining economic phenomena. The results of the experimental papers in economics provide a piece of evidence compatible with the theory that women may have a different attitude toward competition than men, which in turn is one of the plausible theories behind gender wage gaps. Note that the experiment is not discarding other alternative explanations for gender wage gaps, but rather adds an additional possible one. Of course, the important debate on wage gaps is far from being resolved with the results of a particular experiment. A convergence approach in which laboratory results are compared with other sources (e.g., field data, empirical labor market data, etc.) is, in our view, the most beneficial to our understanding of these phenomena.

Economic experiments create an idealized and simplified version of the real world. It is important to understand the limitation of the tool. First, since theory models and experiments operate with simplified versions of reality, it is highly unlikely that they contain all the aspects of reality which translate into precise estimates of how a particular phenomenon occurs. This is why, in our view, calibrating parameters of a particular theoretical model in the laboratory hardly ever makes sense. Both the model and the experiment should provide a sense of direction on the reasons behind a phenomenon and on the possible ways in which different variables interact ("treatment effect").

Experiments should also be kept simple in order for the participants to understand what they are doing. This does not mean that the theoretical explanation motivating the experiment needs to be necessarily simple. Varian (1994) justifies the development of economic theory as a result of insufficient data to explain a phenomenon.

Importantly, in some cases the theoretical model gets too complex for theorists to be able to solve it. In the current state of affairs in the literature, it would be hard to publish an experiment for which we are unable to produce a theoretical benchmark. We argue that this is too restrictive. We should not restrict the discussion only to environments that we can solve using analytical models. By expanding the discussion, we can study empirically environments that are very relevant to real-world economics.

Auctions are a typical example of a market institution for which theoretical analysis, although extensive and insightful, has limitations. When modeling many of the most commonly used auctions formats, there is either no equilibrium or multiplicity of equilibria, and solving the models or selecting equilibria requires very restrictive assumptions. Still, it is important to compare how individuals behave in different auction formats. A classic example is Plot and Smith (1978), who use the laboratory as a "wind-tunnel" in order to compare market institutions. Auction designers and theorists have benefit from the insight developed from market experiments since then.

The dialogue between theory and experiments is ongoing. We can only hope that it will be even more constructive than it was in the past.

NOTES

We are grateful to Tilman Börgers, Ángel Hernando-Veciana, and Steffen Huck for their helpful comments. Pedro Rey-Biel acknowledges financial support from Ministerio de Educacion y Ciencia (ECON2009-0716), Ministerio de Economía y Competitividad (Grant ECO2012-31962), Barcelona GSE Research Network, and the Government of Catalonia (2009SGR-00169).

REFERENCES

Bateman, A. J. 1948. Intra-sexual selection in *Drosophila*. *Heredity* **2**:349–368.

Darwin, C. 1871. *The Descent of Man and Selection in Relation to Sex*. London: John Murray.

Gneezy, U., M. Niederle, and A. Rustichini. 2003. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics* **118**(3):1049–1074.

Niederle, M. 2015. Intelligent Design: The Relationship Between Economic Theory and Experiments: Treatment-Driven Experiments. In *Handbook of Experimental Economic Methodology,* eds. Fréchette, G. R. and A. Schotter. Oxford University Press.

Plot, C. and V. Smith. 1978. An Experimental Examination of Two Exchange Institutions. *Review of Economic Studies* **45**:133–153.

Trivers, R. L. 1972. Parental investment and sexual selection. In *Sexual Selection and the Descent of Man*, ed. B. Campbell. Chicago: Aldine.

Varian, H. L. 1994. What Use Is Economic Theory? *Method and Hist of Econ Thought* 9401001, EconWPA.

PART III

PSYCHOLOGY AND ECONOMICS: A COMPARISON OF METHODS

CHAPTER 9

THE 1-800 CRITIQUE, COUNTEREXAMPLES, AND THE FUTURE OF BEHAVIORAL ECONOMICS

IDO EREV AND BEN GREINER

Introduction

One of the main differences between basic research in psychology and economics involves the trade-off between the "descriptive accuracy" and the "parsimony" of the popular models. Classic papers in psychology tend focus on accuracy; they start with the presentation of new empirical data, and they conclude with the presentation of a new model that provides an accurate summary of these results. For example, consider Sternberg's (1966) classic study of search in short-term memory, which starts with a study of the effect of memory set size on reaction time and concludes with the presentation of a simple model that captures these results.

Classic papers in economics pay more attention to the parsimony and potential generality of the proposed model. For example, consider Akerlof's (1970) classic demonstration of the impossibility of markets for lemons. The paper starts with the observation that rational economic theory implies that markets in which the sellers have more information than the buyers cannot exist. This assertion is not

exactly accurate, but it provided a parsimonious approximation of a number of empirical observations.[1]

The parsimony and generality of traditional economic models of decision making facilitates a broad application. However, in certain settings the predictions of these models just don't match with actual observed behavior. In turn, the main shortcoming of the focus on accuracy in psychological models is captured by the "1-800" critique (Roth, personal communication). According to this critique of mainstream psychological research, psychologists should add a toll-free (1-800) support phone number to their papers and be ready to answer phone calls concerning the predictions of their theories in new settings. The basic argument behind this critique is that the psychological research identified many regularities and described them with as many different models. The boundaries of these models are not always clear, and sometimes they can be used to support contradicting predictions. Thus, it is not clear how these models can be applied.

Behavioral economists try to maintain a third way by building on the attractive features of basic research in both psychology and economics, seeking accuracy while maintaining a clear connection to the general and parsimony model of rational choice. Specifically, the most influential research focuses on two classes of deviations from rational choice. One class involves violations of expected utility theory (von Neumann and Morgenstern, 1947). The clearest counterexample of this type is the Allais paradox (see Allais (1953) and Table 9.1). Prospect theory (see Kahneman and Tversky (1979) and Wakker (2010)), the best-known explanation of this and similar paradoxes, implies that the deviation reflects overweighting of rare events. A second class involves violations of the assumption of self-interest—that is, that people try to maximize only their own outcomes. The clearest counterexamples of this type are observed in the study of a single play of the prisoner dilemma game (see Flood and Dresher (1952) and Table 9.2) and the ultimatum game (see Güth et al. (1982) and Table 9.3). These and similar observations are naturally captured with the assumption of "other regarding preferences" (see Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and Rabin and Charness (2002)).

Table 9.1. The Two Problems used by Kahneman and Tversky (1979) to Replicate Allais' Common Ratio Paradox

	Problem 1		Problem 2
S1:	Choose between: 3000 with certainty	S2:	Choose between: 3000 with probability 0.25 0 otherwise (p = 0.75)
R1:	4000 with probability 0.8 0 otherwise (p = .2)	R2:	4000 with probability 0.2 0 otherwise (p = 0.8)

Note: Rationality (expected utility theory) implies that S1 will be selected in Problem 1 if and only if S2 is selected in Problem 1. The results reveal that most people prefer S1 and R2.

Table 9.2. An Example of a Prisoner Dilemma Game

		Player 2	
		C	D
Player 1	C	1, 1	–1, 2
	D	2, –1	0, 0

Note: Player 1 selects a row, and Player 2 selects a column. The selected cell determines the payoffs. Player 1's payoff is the left entry, and Player 2's payoff is the right entry. Rationality (preferring the dominant strategy) implies D choices in one shot play of this game. Experimental studies (see Rapoport and Chammah (1965)) show nearly 50% C choices.

Table 9.3. A simplified Ultimatum game

Player 1's Choice	Player 2's Choice	Player 1's Payoff	Player 2's Payoff
Up	Up	8	2
	Down	0	0
Down	Up	5	5
	Down	0	0

For example, it is possible that in certain settings some decision makers try to decrease the difference between their own outcome and the outcomes of other agents (inequality aversion).

The focus on counterexamples to rational economic theory can, in theory, solve the 1-800 critique. Once we understand the robust counterexamples, we may be able to refine the rational model to capture the general behavioral tendencies while maintaining parsimony. This solution rests, however, on two nontrivial working assumptions: The first states that the understanding of the robust counterexamples implies the discovery of general behavioral tendencies that drive behavior. The second states that it is possible to capture the counterexamples with a parsimony model that allows clear predictions.

Pesendorfer (2006) questions the descriptive value of the second (clarity) working assumption. In his review of the book *Advances in Behavioral Economics* (Camerer et al., 2004), he states:

> Behavioral economics emphasizes the context-dependence of decision making. A corollary of this observation is that it is difficult to extrapolate from experimental settings to field data or, more generally, economic settings.

> Moreover, not all variables that are shown to matter in some experiment are useful or relevant in economic applications. The question whether a particular variable is useful or even observable for economics rarely comes up in behavioral models, yet the success or failure of modeling innovations often depends on its answer. (Pesendorfer, 2006; p. 720)

In other words, Pesendorfer suggests that the leading behavioral models are not clear. They use concepts that cannot be observed and/or estimated outside the laboratory.

The main goal of the current analysis is to extend Pesendorfer's assertion. It clarifies the shortcomings of the focus on counterexamples, and it reviews recent research that tries to address these shortcomings. The chapter starts with the description of two counter-to-counterexamples—environments in which natural generalizations of the best-known counterexamples to rational economic theory lead to incorrect predictions. The observed behavior deviates from maximization in the opposite direction of the predictions of the popular explanations of the relevant counterexamples. The first "counter-to-counterexample" highlights a tendency to underweight rare events. Thus, it implies a reversal of the pattern captured by prospect theory. The second counter-to-counterexample reflects a deviation from fair and efficient equilibrium. The two counter-to-counterexamples share the same structure: They address famous deviations from rationality that can be reliably observed when experimental subjects respond to a complete description of the incentive structure. The observed behavioral patterns, however, are not general: Reversed patterns emerge when the subjects have to decide based on personal experience.

The chapter is concluded with a review of recent research that tries to address the 1-800 critique by extending the study of counterexamples with a focus on quantitative predictions.

Counter-to-counterexamples

We believe that the most important shortcoming of the focus on counterexamples to rational economic theory is the effect of this focus on the selection of experimental paradigms. In order to discover clear violations of rational economic theory, researchers have to study situations in which this theory leads to clear predictions (that can be rejected). It turns out that the set of situations with this quality is not very large. Many social interactions have multiple equilibria; and when the economic agents rely on personal experience, almost any behavior can be justified as rational under certain assumptions.[2] This observation led behavioral economists to focus on simple "decisions from description": experiments that focus on decisions based on a complete description of the incentive structure. This convention masks the fact that the rationality benchmark is limited and can lead to incorrect generalizations. Two demonstrations of this problem are presented below.

Experience in Individual Decision Making: The Weighting of Rare Events

Kahneman and Tversky (1979) proposed prospect theory to summarize the behavioral regularities documented in the study of individual decisions from description. In all the experiments they considered, the decision makers received a complete description of the incentive structure. Nevertheless, many of the influential applications of prospect theory address situations in which the decision makers are likely to rely on personal experience in the absence of a complete description of the payoff distributions. For example, Benartzi and Thaler (1995) use prospect theory to explain investment decisions, and Camerer et al. (1997) use prospect theory to explain the decisions of taxi drivers.

Recent studies suggest that these and similar applications can lead to incorrect conclusions. Experience does not appear to trigger the behavior captured by prospect theory. There is no evidence for loss aversion in decision from experience (see Erev et al., 2008). Moreover, when decision makers rely on personal experience in binary choice tasks under uncertainty, they tend to deviate from maximization in the direction of underweighting of rare events. This pattern, documented in the study of the behavior of humans (see Barron and Erev (2003) and Hertwig et al. (2004)) and other animals (see Shafir et al. (2008)), is illustrated by the study of the problem presented on the left-hand side of Table 9.4.

Erev et al. (2010) studied this problem under three conditions. Condition Clicking used the "clicking paradigm" described in Table 9.5. The experiment included 100 trials. In each trial the participants were asked to select between two unmarked keys on the computer screen. The left key (option S) yielded a sure gain of 2.7 shekels (1 shekel equaled about 0.2 euro), and the right key (Option R) provided 3.3 shekels in 91% of the trials and a loss of 3.5 shekels in 9% of the trials.

The payoff from the selected key determined the decision maker's payoff for the trial. The decision makers received no prior information concerning the relevant payoff distributions; their information was limited to the presentation of the obtained and forgone payoffs after each trial (thus, exploration was not an issue: the decision makers received feedback concerning the payoff from both

Table 9.4. Choice Task and Choice Rates in Erev et al. (2010)

			The Choice Rate of Option S		
	The Choice Problem	Expected Value	Clicking	Cards	Description
S	2.7 with certainty	2.700	42%	35%	75%
R	3.3 with probability 0.91; –3.5 otherwise	2.688			

Note: The left-hand side presents the basic choice task in Erev et al. (2010). Note that Option S yields a higher expected value at a lower variance of payoffs. Results for the three conditions are presented on the right-hand side.

Table 9.5. The Instructions and Screens in a Study that Uses the Basic Clicking Paradigm

Instructions	Pre Choice	Post Choice
The current experiment includes many trials. Your task, in each trial, is to click on one of the two keys presented on the screen. Each click will result in a payoff that will be presented on the selected key. At the end of the study, one of the trials will be randomly selected, and your payoff in that trial will be added to your show-up fee. Your goal is to maximize your total payoff.	Please select one of the two keys [ ] [ ]	You selected Left. Your payoff is 2.7 [2.7] [3.3]

alternatives after each choice). The payoffs were presented on the keys for one second after each choice.

Note that the "safe" alternative S is associated with higher expected payoffs and lower variance. The proportion of S choices (S-rate) over the 100 trials in condition Clicking was only 42%; the decision makers tended to prefer the riskier, lower expected value, alternative. This deviation from maximization can be captured with the assertion that the decision makers underweight the rare event (the 9% chance to obtain –3.5).

In condition Cards the participants were to select once between two decks of cards. They were told that their payoff will be determined based on a random draw of a single card from the selected deck; the payoff will be the number written on the card. They were allowed to sample the two decks as many times as they wished. One deck corresponded to option S (the number on all the cards was 2.7), and the second deck corresponded to option R (91% of the cards were "3.3" and the rest were "–3.5"). The S rate was 35%.

Finally, condition Description used Kahneman and Tversky's paradigm. The payoff distributions were described to the decision makers. The S rate under this condition was 75%. The study of additional problems reveals that the difference between the three conditions does not reflect the higher maximization rate in the Description condition. The results are best summarized with the assertion of underweighting of rare events in the two experience conditions (Clicking and Cards) and overweighting of rare events in the Description condition.

Experience in Social Conflicts: Efficiency, Fairness, and Search

Basic research of social interactions highlights robust violations of the assumption that people try to maximize their own outcomes. The clearest counterexamples of this type were observed in the study of a single play of the Prisoner Dilemma game

and the Ultimatum Game described in Tables 9.1 to 9.3. These interesting observations imply a deviation from rational choice in the direction of efficient and fair outcomes.

The classic demonstrations of these counterexamples involve decisions from descriptions: The participants receive a precise description of the incentive structure. Can we generalize the results to situations with limited prior information concerning the incentive structure? This question has received limited attention; yet, the obtained results are interesting. For example, Mitzkewitz and Nagel (1993; see a clarification in Rapoport and Sundali (1996)) found that a slight constraint on the information available in the Ultimatum game can reduce the importance of other-regarding preferences. Recall that the game includes two stages. In the first stage, one player—the proposer—proposes a division of a pie between herself and a second player. In the second stage, the second player—the responder—can accept or reject the proposal. In the original game, the size of the pie is known to both players. Mitzkewitz and Nagel (1993) compared this condition to a variant in which only the proposer knows the size; the receiver's information is limited to the distribution of the possible values. For example, the proposer knows that the size is 10, and the receiver knows that it is between 1 and 10. The results reveal that the lack of complete information moved behavior toward the rational (subgame-perfect equilibrium) prediction: It reduced the proposal and increased the acceptance rate of a given proposal.

Under one explanation of this effect, the availability of full information leads many decision makers to focus on fair and efficient outcomes (as suggested by Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Rabin and Charness (2002); see also the review by Cooper and Kagel (forthcoming)). When the information is incomplete and the game is played repeatedly, behavior is driven by exploration. And when the information does not allow evaluation of fairness and efficiency, these factors are not likely to affect behavior. In order to clarify the implication of this hypothesis, it is constructive to consider the 5×5 a symmetric Stag Hunt game presented in Table 9.6. Note that the game has two equilibrium points. The E/E equilibrium is efficient (payoff dominant) and fair: Both players win 12 (joint payoff of 24) under this equilibrium. The A/A equilibrium is inefficient (joint payoff of 15)

Table 9.6. A 5 × 5 Asymmetric Stag Hunt Game

		Player 2				
		A	B	C	D	E
Player 1	A	10, 5	9, 0	9, 0	9, 0	9, 0
	B	0, 4	0, 0	0, 0	0, 0	0, 0
	C	0, 4	0, 0	0, 0	0, 0	0, 0
	D	0, 4	0, 0	0, 0	0, 0	0, 0
	E	0, 4	0, 0	0, 0	0, 0	12, 12

and unfair (one player wins 10, and the second wins 5), but it is the risk dominant equilibrium. Assuming the game is played repeatedly with fixed matching, the current logic implies a large effect of the availability of prior information: With a complete description of the game, most pairs are expected to converge to the fair and efficient equilibrium (E/E). However, when the prior information is limited, many pairs are likely to converge to the unfair and inefficient risk dominant equilibrium (A/A).

We tested this hypothesis experimentally. Twenty-four pairs of subjects played the game in Table 9.6 for 50 trials (using fixed pair matching). For each pair, the location of the A/A and E/E equilibria was determined randomly before round 1. Each pair played the game under one of two conditions, "Description" or "Experience." The participants received a complete prior description of the game in condition Description, but no prior information on the game payoffs in condition Experience. In both conditions, the feedback after each trial was limited to own obtained outcomes.

The results, summarized in Figure 9.1, reveal a clear effect of the prior information. The proportion of fair and efficient outcome (E/E) in the last 10 trials was 84% in condition Description and only 25% in condition Experience. The proportion of the risk dominant equilibrium outcome (A/A) was 16% in condition Description and 59% in condition Experience.

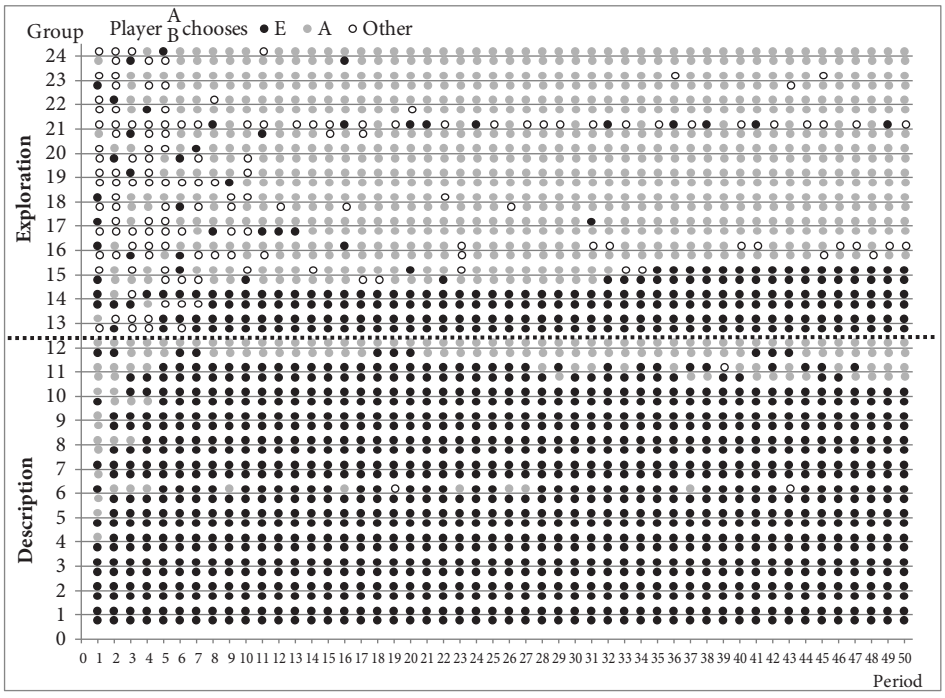


Figure 9.1. Choices of the two players over 50 periods.

Table 9.7. A 5 × 5 Coordination Game with a Unique Fair and Efficient Equilibrium

		Player 2				
		A	B	C	D	E
Player 1	A	10, 5	9, 0	9, 0	9, 0	9, G
	B	0, 4	0, 0	0, 0	0, 0	0, 0
	C	0, 4	0, 0	0, 0	0, 0	0, 0
	D	0, 4	0, 0	0, 0	0, 0	0, 0
	E	0, 4	0, 0	0, 0	0, 0	12, 12

Note: The outcome G (for Player 2 in the cell (A, E)) is a realization of a gamble that pays 1000 with probability 0.01, and 0 otherwise.

Consider now the coordination game presented in Table 9.7. This game is identical to the Stag Hunt game in Table 9.6 with one exception: Player 2's payoff in the upper-right cell was changed from "0" to a gamble that pays "1000 with probability 0.01, and 0 otherwise." This change eliminates the risk dominant equilibrium and creates a coordination game with a unique, fair, and efficient equilibrium. It is easy to see, however, that this change is not likely to change behavior. With complete information, participants are still likely to prefer the fair and efficient outcome, and in decisions from experience the participants are still expected to converge to the A/A cell that implies the unfair and inefficient outcome. This cell does not establish an equilibrium anymore, but it is hard to think about a learning process that would not move behavior toward that point.

In summary, the current analysis suggests that social interactions that evolve from experience can lead to very different patterns of behavior than the ones documented in mainstream research, which focuses on decisions based on a complete description of the game. Experimental studies of decisions from description highlight deviations from rational choice equilibrium in the direction of fair and efficient outcomes. Decision making relying on experience can exhibit the opposite pattern. Whereas the "decisions from experience" results do not imply deviation from rationality, they can be important. It is possible that they capture a frequent and important situation. Indeed, it is possible that many social conflicts (including marital and national conflicts) are the product of tedious exploration problems, rather than deep incentives and/or emotions. We do not know how many natural conflicts reflect exploration failure, but it seems safe to assert that the focus on rational choice and counterexamples to rational decision theory is not likely to shed light on this important issue.

Quantitative Predictions

The results summarized above suggest that it is important to extend behavioral economic research beyond the focus on counterexamples. It is important to quantify

the relevant models and clarify their boundaries; and it is important to advance the experimental analysis beyond the study of decisions from description. These extensions should lead to the development of quantitative models of choice behavior that capture the famous counterexamples to rational decision theory, and they should allow useful predictions even when the rational prescriptions are ambiguous.

The idea that quantitative models can be useful is, of course, not new and/or controversial. Nevertheless, it seems that behavioral economists tend to avoid quantitative analyses. The most influential studies can be described as first steps toward quantitative analyses. That is, they present models that can be quantified, and they leave the task of testing the model's predictive value to future research. Erev et al. (2010) assert that one contributor to this observation is a problematic incentive structure for behavioral researchers: The evaluation of quantitative predictions tends to be more expensive and less interesting than the study of counterexamples. Research that leaves the evaluation of the quantitative predictions to future studies can start with a presentation of a few interesting phenomena and conclude with the presentation of a potentially general and insightful model that captures them. To study quantitative predictions, on the other hand, researchers have to consider a wide set of problems, and wide sets tend to be less interesting than the best example. The researcher then has to estimate models and has to run another large and boring (random sampling can help) study in order to compare the different models. Moreover, the cost of the effort increases by the likely rejection of the target model (i.e., if the sample is large enough, all models are expected to be rejected).

Note that this incentive structure can be problematic in two ways. First, it might reflect a public good problem: The research field would benefit from careful study of quantitative predictions, but each researcher is motivated to study more rewarding problems. Under a second interpretation, behavioral economists exhibit underweighting of rare events in decisions from experience (like the participants in the clicking paradigm described above). That is, while the development of quantitative models will most likely lead to boring outcomes, there is a small probability that it will lead to the discovery of very important and robust regularities, which, however, is underweighted. At any event, a modification of the incentive structure in producing new behavioral evidence is desirable.

One procedure that has the potential to overcome this incentive structure is presented in Erev et al. (2010; also see a similar idea in Arifovic et al. (2006)). The first three coauthors of that paper tried to increase the attractiveness of the study of quantitative predictions by the organization of three open choice prediction competitions that can address the high cost and boredom problems. They run the necessary boring studies, and they challenged other researchers to predict the results. All three competitions focused on the prediction of binary choices between a safe prospect that provides a medium payoff (referred to as M) with certainty and a risky prospect that yields a high payoff (H) with probability Ph and a low payoff (L) otherwise. Thus, the basic choice problem is:

Safe: M with certainty
Risky: H with probability Ph; L otherwise (with probability 1-Ph)

The four parameters (M, H, Ph, and L) were randomly selected with a well-defined algorithm that implies the following: (1) The possible payoffs were between –30 and +30 shekels (1 shekel equaled about $0.3); (2) L < H; (3) M was between L and H in 95% of the problems; and (4) the difference between the expected values of the two prospects was relatively small.

Each competition focused on a distinct experimental condition, with the objective being to predict the behavior of the experimental participants in that condition. The three conditions include: Description, Cards, and Clicking. In condition "Description," the experimental participants were asked to make a single choice based on a description of the prospects (as in the decisions under risk paradigm considered by Allais (1953) and Kahneman and Tversky (1979)). In the condition "Cards" (Experience-sampling), participants made one-shot decisions from experience (as in Hertwig et al. (2004)): They were asked to select once between two decks of cards that were represented by two unmarked keys. The composition of card decks was determined by the two prospects, though the participants were not informed about the distributions. Instead, they had to base their decisions on their personal experience. Specifically, the participants were allowed to sample (with replacement) each of the decks as many times as they wished before moving to the choice stage. The choice stage determined the trial's payoff. During this stage they were asked to choose once between the two decks. In the condition "Clicking" (Experience-repeated), participants made (100) repeated decisions from experience (as in Barron and Erev (2003)). The two prospects were represented by two unmarked keys. Each choice was followed by a presentation of the obtained payoff.

Each of the three competitions was based on the data from two experimental sessions: an estimation session and a competition session. The two sessions for each condition used the same method and examined similar, but not identical, decision problems and decision makers as described below. The estimation sessions were run in March 2008. After the completion of these experimental sessions, the organizers posted the data and several baseline models on the web, and they challenged researchers to participate in three competitions that focus on the prediction of the data of the second (competition) set of sessions.[3]

Researchers participating in the competitions were allowed to study the results of the estimation study. Their goal was to develop a model that would predict the results (the mean choice proportion over all choices in each problem) of the competition study. The model had to be implemented in a computer program that reads the payoff distributions of the relevant gambles as an input and predicts the proportion of risky choices as an output. The submitted models were ranked based on the mean squared deviation (MSD) between the predicted and the observed choice proportions. The main advantage of this measure is its relationship with traditional

statistics (like regression, *t*-test and the *d*-statistic) and its intuitive interpretation. These attractive features are clarified with the computation of the ENO (equivalent number of observations) order-maintaining transformation of the MSD scores (see Erev et al. (2007)). The ENO of a model is an estimation of the size of the experiment that has to be run to obtain predictions that are more accurate than the model's prediction. For example, if a model's prediction of the probability of risky choices in a particular problem has an ENO of 10, this prediction is expected to be as accurate as the prediction based on the observed proportion of risky choices in an experimental study of that problem with 10 participants.

Each problem was faced by 20 participants (Technion students) in a computerized setting. The participants received 25 shekels show-up fee, and the payoff from their selected prospect in one randomly selected trial.

Twenty-three models were submitted to participate in the different competitions; eight to the Description condition, seven to the E-sampling condition, and eight to the E-repeated condition. The submitted models involved a large span of methods that include logistic regression, ACT-R-based cognitive modeling, neural networks, production rules, and basic mathematical models.

Evaluation of the results reveals three interesting observations. First, the raw data highlight (a) high correspondence (correlation above 0.8) between the two Experience conditions and (b) a negative correlation between these conditions and the Description condition. Analysis of this difference reveals that it is driven by the effect of rare events. The proportion of risky choices increased with Ph in the two Experience conditions and decreased with Ph in the Description condition. This pattern emphasizes the robustness of the assertion that decision makers behave as if they underweight rare events in the Experience conditions and overweight rare events in the Description condition (see Barron and Erev (2003)).

Examination of the submitted models and their predictive value complements the first observation. Very different models were submitted for, and won, the three competitions. The best models in the "decisions from description" condition were stochastic variants of prospect theory. They shared the assumption of overweighting of rare events. The best models in the two "decisions from experience" conditions shared the assumption that decision makers rely on small samples of experiences. This assumption implies a tendency to underweight rare events.

A third interesting observation comes from a comparison of the predictive value of the different models as measured by the models' ENO. This analysis shows that models that were proposed to capture counterexamples have a low ENO value. For example, the ENO of the original version of prospect theory is around 2. However, a minimal modification of this model, the addition of a stochastic response rule, dramatically increases its predictive value. The ENO of the best stochastic variant of prospect theory is 81. This and similar results in the experience conditions (the best models in conditions E-Sampling and E-Repeated had an ENO of 35 and 47, respectively) suggest that behavioral models can do much better than fitting the data: They may provide very useful predictions.[4]

Summary

The 1-800 critique is one of the most important challenges to the hope of using laboratory research to predict social behavior in natural settings. It captures the fact that different experiments reveal different behavioral regularities and that it is typically difficult to know which of the different regularities is more important in a particular application. Mainstream research in behavioral economics tries to address this critique by using the parsimonious rational decision theory as a benchmark. Most experimental studies focus on counterexamples to this theory, and the leading descriptive models are generalizations of parsimonious rational models. The results summarized above question the value of this approach. The leading behavioral models cannot be easily applied without a support call center. Moreover, in certain settings, reasonable attempts to apply these models lead to incorrect predictions. The observed behavior deviates from maximization in the opposite direction of the predictions of the most popular explanations of the classical counterexamples.

Part of the problem appears to reflect a shortcoming of the focus of rational economic theory. This theory is parsimonious and potentially general, but the set of situations for which it provides clear predictions is limited. Thus, there are many situations in which the reliance on this theory as a benchmark cannot support clear predictions of behavior.

We believe that these problems can be addressed with a focus on quantitative predictions. The development of clear models that provide useful quantitative predictions of choice behavior in well-defined spaces of situations can solve the 1-800 problem. The value of this approach is expected to increase with its popularity. The accumulation of research will improve the models and increase the spaces of studied situations. The choice prediction competition procedure, used by Erev et al. (2010) and described above, presents one way to facilitate and incentivize this research.

NOTES

Part of this research was conducted when the authors were at Harvard Business School. We thank Alvin E. Roth and Peter Wakker for their useful comments.

1. Moreover, the publication of this influential paper facilitated "lemon laws" (like the Magnuson–Moss Warranty Act) that have reduced the likelihood of markets for lemons.

2. This observation implies a difference between generality and clear predictions. Rational economic theory is general; unlike behavioral models, rational models do not assume context dependence. Nevertheless, this does not imply that it can be applied in a context-independent fashion. In most cases the prediction of behavior, even under the rationality assumption, requires context-dependent information (or assumptions).

3. The main prize for the winners was an invitation to coauthor the paper that describes the competitions.

4. Two observations clarify the meaning of these ENO values. First, 20 years ago, Artillery officers were instructed to use Newton laws (to predict the behavior of a cannon) under the assumption that the laws' ENO is between 2 and 3. That is, they were supposed to use Newton laws for the first three shots, and then rely on the mean of the experiment (the mean of the first three falls) to predict the fourth. Second, Erev et al. (2007) show that $t^2 = (n/ENO) + 1$, where t is the t-statistic of the hypothesis that the model is "accurate" (the model prediction is the null hypothesis) and n is the number of experimental observations. Thus, an ENO of 81 means that one has to run more than 300 observations in order to reject the model at the .05 level (obtain an absolute t-statistic larger than 1.96).

REFERENCES

Akerlof, G. A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* **84**:488–500.

Allais, M. 1953. Le Comportement de l'Homme Rationnel Devant le Risque, Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica* **21**:503–546.

Arifovic, J., R. D. McKelvey, and S. Pevnitskaya. 2006. An Initial Implementation of the Turing Tournament to Learning in Repeated Two-Person Games. *Games and Economic Behavior* **57**:93–122.

Barron, G. and I. Erev. 2003. Small Feedback-Based Decisions and Their Limited Correspondence to Description Based Decisions. *Journal of Behavioral Decision Making* **16**:215–233.

Benartzi, S. and R. A. Thaler. 1995. Myopic Loss Aversion and the Equity Premium Puzzle. *Quarterly Journal of Economics* **110**:73–92.

Bolton, G. E. and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* **90**:166–193.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler. 1997. Labor Supply of New York City Cabdrivers: One Day at a Time. *Quarterly Journal of Economics* **112**:408–441.

Camerer, C. F., G. Loewenstein, and M. Rabin (eds.). 2004. *Advances in Behavioral Economics*. New York: Princeton University Press.

Cooper, D. J. and J. Kagel. Forthcoming. Other-regarding Preferences. In *The Handbook of Experimental Economics*, Volume 2, eds. J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.

Erev, I. and A. E. Roth. 1998. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review* **88**:848–881.

Erev, I., A. E. Roth, R. L. Slonim, and G. Barron. 2007. Learning and Equilibrium as Useful Approximations: Accuracy of Prediction on Randomly Selected Constant Sum Games. *Economic Theory* **33**:29–51.

Erev, I., E. Ert, and E. Yechiam. 2008. Loss Aversion, Diminishing Sensitivity, and the Effect of Experience on Repeated Decisions. *Journal of Behavioral Decision Making* **21**:575–597.

Erev, I., E. Ert, A. E. Roth, E. Haruvy, S. Herzog, R. Hau, R. Hertwig, T. Stewart, R. West, and C. Lebiere. 2010. A Choice Prediction Competition, for Choices from Experience and from Description. *Journal of Behavioral Decision Making* **23**:15–57.

Fehr, E. and K. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* **114**:817–868.
Flood, M. and M. Dresher. 1952. Some Experimental Games. *Research memorandum RM-789*. Santa Monica, CA: RAND Corporation.
Güth, W., R. Schmittberger, and B. Schwarze. 1982. An Experimental Analysis of Ultimatum Games. *Journal of Economic Behavior and Organization* **3**:367–388.
Hertwig, R., G. Barron, E. U. Weber, and I. Erev. 2004. Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science* **15**:534–539.
Kahneman, D. and A. Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* **47**:263–291.
Mitzkewitz, M. and R. Nagel 1993. Experimental Results on Ultimatum Games with Incomplete Information. *International Journal of Game Theory* **22**:171–198.
Pesendorfer, W. 2006. Behavioral Economics Comes of Age: A Review Essay on "Advances in Behavioral Economics." *Journal of Economic Literature* **44**:712–721.
Rabin, M. and G. Charness. 2002. Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* **117**:817–869.
Rapoport, A. and A. M. Chammah. 1965. *Prisoner's Dilemma*. Ann Arbor, MI: The University of Michigan Press.
Rapoport, A. and J. Sundali. 1996. Ultimatums in Two Person Bargaining with One Sided Uncertainty: Offer Games. *International Journal of Game Theory* **25**:475–494.
Shafir, S., T. Reich, E. Tsur, I. Erev, and A. Lotem. 2008. Perceptual Accuracy and Conflicting Effects of Certainty on Risk-Taking Behavior. *Nature* **453**:917–921.
Sternberg, S. 1966. High-Speed Scanning in Human Memory. *Science* **153**:652–654.
Tversky, A. and D. Kahneman. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* **5**:297–323.
Wakker, P. P. 2010. *Prospect Theory for Risk and Ambiguity*. Cambridge, UK: Cambridge University Press.
von Neumannvon Neumann, J. and O. Morgenstern. 1947. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University.

CHAPTER 10

A GENERAL MODEL FOR EXPERIMENTAL INQUIRY IN ECONOMICS AND SOCIAL PSYCHOLOGY

J. KEITH MURNIGHAN

Introduction

This chapter focuses on the intersection of economics, social psychology, and the experimental method. No other truly social science uses experimental methods as much as these two fields. Indeed, as Thomas Ross and I have argued before (Murnighan and Ross, 1999), experiments provide a common arena for these two important fields to effectively collaborate. Unfortunately, increasingly strident philosophical stances on what is and is not appropriate methodologically have limited the collaborative potential of these two related fields.

I start this chapter by presenting some definitions and some historical perspectives on the two fields. I then proceed to focus on concepts that are not just common but central to both modes of inquiry. I then provide a methodological framework that might satisfy the desires and goals of both fields while simultaneously expanding opportunities not only for each field to contribute to the other but also to expand what we know about the phenomena that we study.

Economics and Social Psychology—Definitions

As a non-economist searching for a single definition of the field of economics, I was both perplexed and dismayed by a bewildering array of definitions. No single, accepted definition of economics emerged from my search of the literature. Elements that were repeated in the various definitions included prices, production, spending, income, individual and social action, maximizing, markets, equilibria, preferences, satisfaction, self-interest, choices, goals, scarcity, wealth, and human relationships. A singular definition that could achieve the support of a consensus of experts seems unimaginable. At the same time, as a Supreme Court Judge said about another difficult-to-define concept, it is not difficult to conclude that most people will recognize economics when they see it.

Definitions of social psychology seem more contained, possibly because it is a subfield of the more general field of psychology. William James (1890) defined the broader field of psychology as "the description and explanation of states of consciousness in human beings . . . the study of . . . sensations, desires, emotions, cognitions, reasonings, decision, volitions, and the like." Social psychology refines this definition by focusing on three critical concepts of individual psychological experience—cognitions, affect, and behavior—within an explicitly social context, that is, "the actual, imagined, or implied presence of others" (Allport, 1985).

All of the central, definitional elements in the definition of social psychology are critical in economics. At the same time, economics often takes a broader perspective, focusing not just on individual economic actors but on the amalgamation of individuals' choices and behaviors. As Ross and I noted in our earlier observations, "Microeconomics studies the choices and behaviors of individuals and firms; social psychology studies the thoughts, behaviors, and actions of individuals and groups in social contexts" (Murnighan and Ross, 1999, p. 1).

Economics and Social Psychology—Historical Perspectives

The fields of psychology and economics both pride themselves on being able to make advances in understanding human behavior. Both fields have congratulated themselves on the progress that they have made in challenging endeavors. B. F. Skinner, for example, in *Science and Human Behavior* (Skinner, 1953, p. 15), noted that "Behavior is a difficult subject matter, not because it is inaccessible, but because it is *extremely complex*." Similarly, Paul Samuelson, in the first edition of his *Economics* (Samuelson, 1948), said that "It is the first task of modern economic science to describe, to analyze, to explain, to correlate these fluctuations of national

income. Both boom and slump, price inflation and deflation, are our concern. *This is a difficult and complicated task*. Because of the complexity of human and social behavior, we cannot hope to attain the precision of a few of the physical sciences. We cannot perform the controlled experiments of the chemist or biologist."

These are just two viewpoints, but an informal scan suggests that inquiry in the social sciences is indeed a complex, challenging endeavor. The two fields also have lofty goals. Another (lengthy) quote from Murnighan and Ross (1999):

> The objective of much of social psychology is to better understand how individuals make decisions in social situations. The focus then is on what motivations drive individuals to make the decisions they make—and the goal is explaining every individual's decisions. Economics, on the other hand, is ultimately about explaining aggregates like market prices and quantities, incomes, employment and market efficiency. The goal is to explain collective behavior rather than individual behavior. To do this, economists have had to construct models of individual behavior *as instruments* for helping them explain the aggregates about which they really care. As a result, economists are prepared to sacrifice success at explaining every individual's choices if, as a result, they can build a better (i.e. more accurate, more tractable, more testable) model of collective, market behavior. . . . Social psychologists, on the other hand, may try to explain behavior that arises only in select circumstances.

In other words, the goals of the two fields differ, but the focus of their attention, ultimately, involves individuals' choices—aggregated for economically important reasons, isolated for psychologically important reasons. The completely common, identical meeting place for both fields, then, is the individual choice. In this chapter I assume that both fields want to improve the methods that they can use to observe and understand these choices. The challenge is to overcome the notion that "the two fields promote different kinds of thinking and different philosophies, and these differences make it difficult for people in the two disciplines to collaborate, much less appreciate each others' work" (Murnighan and Ross, 1999).

Methods and Approaches

The methods of experimental economics have developed rapidly. Although economic scientists approach experiments in many of the same ways that social psychologists do, they differ in some fundamental ways as well. The two that have received the most attention, incentives and deception, have, from a social psychologist's point of view, led many economists to reject much of the work in social psychology. This is truly unfortunate.

In this chapter I will not dwell on these two topics as they have been addressed at length in other work (e.g., Croson, 2006; Jamison et al., 2007). The crux of the two fields' disagreements on incentives is the requirement in economic research

that people be paid relatively large sums of money that are contingent on their actions and the results of their actions. The puzzling part of this strong requirement is that considerable research (e.g., Camerer and Thaler, 1995) has shown that, in many paradigms, changes in monetary incentives do not change the central tendencies of the findings. Instead, the most common effect of low incentives seems to be an increase in error variance. (There are certainly exceptions, particularly when incentives are considerable.)

The disagreement about deception is more fundamental and more heated, so much so that many economists treat deception as if it is inherently immoral. This is surprising as the issue can also be framed as being strictly empirical: Economists are concerned that deception in one experiment will bias responses or participation in subsequent experiments. They decry the use of deception of any kind and claim that any researcher who uses deception poisons the pond for all future experimenters.

It is to the field of economics' great credit that some of its researchers have investigated their empirical concerns, both for incentives (with the results noted above) and for deception. Jamison et al. (2007), for instance, conducted a careful experiment in which some participants were deceived and others were not and the participants' future behavior could be assessed. They found selected, nonsignificant effects of prior deception on volunteering rates and on behavior in a subsequent experiment. They note that researchers in economics think that their findings support the validity of their concerns but that researchers in social psychology think that their findings indicate that economists' concerns are essentially groundless. Thus, the empirical findings to date do not eliminate the conflict, or the philosophical differences that have become associated with it. Rather than pursuing these issues further, the rest of this chapter addresses the advantages of the two fields' approaches to experiments.

Economics' Advantages

The primary advantage of an economic approach to experiments is the strength of its theory: The formal mathematics of economic theory means that its behavioral predictions are particularly sharp. Once a situation satisfies a set of basic assumptions, economics can make truly specific predictions about how rational individuals should behave. The theories also leave considerable room for psychological input, as psychologists are interested in how people actually do behave. For psychologists, for instance, there is no need to assume strict rationality. This is not to say that we don't normally expect people to be consistent, but personal consistency and economic rationality may not be equivalent in psychological conceptions of choice.

A second advantage offered by the field of experimental economics, and this may be my own idiosyncratic view, consists of the creativity of its games: Economists have introduced many of the games that psychologists love to use

in their experiments, including the prisoners' dilemma, the ultimatum game, the dictator game, the trust/investment game, and the like. These are beautifully constructed, compelling situations that have the potential of surfacing a wide array of interesting individual and interpersonal actions, reactions, and interactions.

A sticking point for both economists and psychologists who study experimental games is that our participants have a nasty habit of transforming the games that we present to them. In other words, while we may present a clean, clear set of rules and outcomes, participants may view the game in ways that differ from the way that we view them.

Kelley and Thibaut (1978) were the first to make this clear. As they put it, "Raw matrix values are simply not satisfactory predictors of behavior" (p. 15). They go on to observe: "In effect, by responding to aspects of pattern in the *given* matrix, the actors *transform* it into a new matrix, the *effective* one, which is then closely linked to their behavior" (p. 17). Some of our own recent research (Zhong et al., 2007) provides a pertinent example of this process in the context of the prisoners' dilemma. We found that, while researchers have often interpreted the choices in a prisoners' dilemma as cooperation or defection (e.g., Rapoport and Chammah, 1965), participants appear to be much more focused on the simple analytics of the numbers in the matrices and don't necessarily conceptualize their choices in the same way that researchers do. When game choices were given verbal labels, as cooperation and defection, for instance, participants were significantly more cooperative than they were when their choices were not labeled—that is, when they were presented without verbal labels, as has happened in most prior research.

The conclusion here is simple. As Berg et al. (1995) also noted in their first presentation of the Investment Game, people who did not reciprocate did not necessarily interpreted their counterparts' trusting choices as evidence of trust. Thus, all of us, economists and social psychologists alike, must be particularly careful that we are interpreting our participants' choices in our experiments in the same way that our participants are interpreting them.

Social Psychology's Advantages

Social psychology's attention to socially charged situations and its relatively consistent avoidance of formal mathematics has led to theoretical models that are not as well specified as economics'. At the same time, social psychology preceded economics in its use of experiments as a major mode of inquiry. Indeed, its refinement of the experimental method may be a greater accomplishment for the field than its theoretical developments.

Social psychology, like any field, has gone through a variety of phases, and the value of different aspects of experimental methodology have ebbed and flowed within these phases. In this chapter I attempt to take a long-range view of social psychology's experimental methods, advocating the best approaches taken in its

different phases. In so doing, I will focus on two central aspects of social psychological research, reliability and internal validity, and two of its corollaries: conceptual replication and the identification of underlying psychological mechanisms. (I will explicitly consider external validity as important but secondary, as this seems to be a valuable but not necessary component of the theory-testing research in both economics and social psychology.)

Like experimental economics, much of the experimental work in social psychology tests theoretical propositions and implications. Staw's (1976) early work on the escalation of commitment, for instance, pitted the predictions of reinforcement theory against his own model, based primarily on self-justification, that people might not always refrain from or reduce their future investments after their initial choices had led to negative outcomes. Theory-testing research is well served when it follows the recommendations of John Platt (Platt, 1964), who articulated beautifully the benefits of strong inference—that is, the pitting of several theories' predictions against one another. (Staw's work is a fine example of the fruitfulness of this approach.)

Strong inference is a particularly apt strategy for experimental economics because of the wealth of specific economic models. This approach to research has a natural affinity with strong theory. Yet experimental economics does not use strong inference as a normal, standard approach to its theory-testing studies.

Social psychology does not claim that its theories are strong or specific. Thus, strong inference has its place in social psychology, as does theory testing in general. But social psychological research also takes a different approach, not necessarily to test theory but instead to demonstrate important behavioral regularities that are, on their face, prior to the research, counterintuitive. Work on decision biases like the hot hand in basketball (Gilovich et al., 1985) or prospect theory (Kahneman and Tversky, 1971) or the anchoring and insufficient adjustment effect (e.g., Ku et al., 2006) are examples of research endeavors that are essentially *demonstration studies*: The goal of the research is to demonstrate, repeatedly and in different ways, a reliable effect that is counterintuitive. In the decision biases research, for instance, researchers have sought instances when individuals depart from the kinds of rational, consistent judgments that might otherwise seem to serve their best interests. The accumulation of a series of independent research projects then begins to map the domain of the phenomenon, ultimately defining its boundary conditions.

The hot hand paper provides an excellent example of this approach. The paper first presented findings indicating that people believed that the shots taken by basketball players who had made several consecutive previous shots would have a higher probability of subsequently successful shots than those taken by basketball players who had not made several shots in a row. The second study in the paper analyzed the shooting performance of nine members of the Philadelphia 76ers basketball team in the 1980–1981 season. These data clearly showed that the probability of making a shot after have made one, two or three previous shots in succession was no higher than the probability of making a shot after having *missed* one, two,

or three previous shots in succession. Thus, study 2's analyses disputed individual observers' common beliefs. The data also showed that, for a player who hit 50% of his shots, on average, the number of consecutive sequences of scoring shots did not depart from the number of consecutive sequences of heads one would expect to observe in a string of coin tosses. This analysis suggested that there was nothing even warm about a hot hand; instead, the result of a basketball players' shot could not likely be predicted any more accurately than the results of a coin toss. A third study showed that a subset of the same Philadelphia 76ers, plus their coach, also believed in the existence of the hot hand, even though their own behavior did not support it. Additional analyses of the shooting performance of two other professional teams continued to find the same absence of any evidence of a hot hand.

Their fourth study assessed the impact of a hot hand on free-throw shooting; these data countered the possibility that game conditions might have changed to influence shooting success. These results, however, conceptually replicated the earlier findings, showing no evidence for a hot hand. Finally, the researchers conducted an experiment with college basketball players and a set of independent observers to determine whether the players themselves or "objective," self-interested observers might be able to predict, in advance, the likelihood of a player making a shot. Even though both groups could choose higher stakes when they felt confident about an upcoming shot and lower stakes when they felt that their hands were not hot, neither the players nor the observers were successful in their predictions, and they also made no money from their bets. Thus, in a predictive sense, neither the observers, all of whom were avid basketball fans, nor the players themselves could accurately identify the existence of a hot hand, even though, to this day, people seem to feel that it exists and affects the outcomes of a competition.

In fact, basketball fans and players and researchers have vehemently disputed these findings (Tversky and Gilovich, 1989). Yet both the approach and the findings of this research were clear: They identified what might appear to be commonly understood and demonstrated that is not so common and not so well understood.

In this case, the research did not delve into the underlying psychological mechanisms for these common misunderstandings. This, however, has become a common approach in recent social psychological research. To take one example, Gillian Ku, Adam Galinsky, and I investigated the effects of the starting price of an item as a potential anchor for bidders on the final selling price of that item at auction. The anchoring-and-insufficient-adjustment bias (Tversky and Kahneman, 1974) in auctions focuses on the possibility that starting price might signal value, suggesting that a high starting price will likely lead to a high final price. In our research, we noted that prior work on anchoring had investigated the insufficient adjustment process in individuals but that, in auctions, the process is more socially determined because many individuals can bid. Thus, we showed that, counter to normal anchoring effects, in auctions a low starting price often led to a higher final price than a high starting price.

Our paper included a series of six studies to show the underlying reasons for this observation. The studies included three scenario-based experiments (using Shubik's dollar auction, 1971; an eBay shirt auction simulation done individually; and another eBay simulation, this time done collectively, for a Caribbean vacation), in which participants were given a scenario and indicated what they would do if they actually faced that scenario, and three sets of archival analyses of previous eBay auctions—of Tabriz rugs (probably privately valued items), Nikon cameras (probably common-valued items), and low- and high-starting-priced Tommy Bahama shirts by the same seller. The last of these archival analyses was a field experiment (even though we did not act as the seller; we were just fortunate that an independent seller varies his starting prices).

Our findings documented the impact of three underlying processes for the reversal of the normal anchoring effect: Lower starting prices reduced barriers to entry, thereby increasing traffic and, ultimately, final prices; lower starting prices enticed bidders to invest time and energy, essentially creating sunk costs and, consequently, escalating their commitment to purchasing the item; and, finally, the traffic that resulted from lower starting prices led other bidders to infer value in the item, thereby increasing traffic further as well as final prices. A final study in the paper showed that a barrier to entry that limited traffic—instances of misspelling Michael Jordan's name in the title of an eBay listing—reduced traffic and, in this instance, rather than in comparable instances of correct spellings of Michael Jordan's name in the title, starting prices were positively rather than negatively correlated with final prices.

This research investigated an economic phenomenon—pricing in auctions—from a decidedly social psychological point of view. We found that psychological and social factors, as well as what might be considered an economic factor, barriers to entry, influenced the final outcomes of these auctions. As a result, we identified new boundary conditions to the anchoring-and-insufficient-adjustment bias and demonstrated how, when, and why social processes can interrupt a well-established individual decision process.

From a methodological perspective, these examples identify the value of multiple studies that provide conceptual replications of the same underlying hypothesis. By demonstrating an effect repeatedly in different contexts with different operationalizations of the same independent and dependent variables, we can have greater confidence in the underlying conceptually based hypothesis. When researchers use only single experiments, it is difficult to conclude that the results pertain to the general hypothesis or simply to the current operationalizations of the variables in the hypothesis. When the research is theoretically driven and the variables satisfy the assumptions of the theory, the findings become more potent. At the same time, abstract concepts can all be operationalized in multiple ways, and showing a consistently significant relationship among multiple operationalizations of the same abstract concepts provides a stronger test of any hypothesis.

One side note may also be worth emphasizing. Research in both economics and social psychology has neglected the study of emotions. Both fields have focused productively on the impact of the behaviors and cognitions of individuals, groups, and larger social entities. The study of emotions has been more difficult and has naturally led researchers to pursue the path of least resistance—that is, the easier questions that are more tractable, rather than trying to identify the impact of what are often fleeting emotions. In some of our work on the influence of emotions in trusting and social dilemma decisions (Lount and Murnighan, 2006), for instance, we have found that measuring emotions presents reliability challenges, as people's emotions do not always persist for long, and that manipulating emotions—by showing videos, for example, of Robin Williams as Mrs. Doubtfire to stimulate happiness—are relatively heavy-handed and also only limitedly productive in terms of the time that these emotions are sustained. Nevertheless, as we approach and study individuals who are making important strategic decisions, it should be hard to ignore the importance of emotions as an integral part of this process.

Social psychology's ongoing refinements of its experimental methods consistently address its basic, elementary targets of inquiry: behavior, cognitions, and emotions. It is actually difficult to imagine aspects of human action that do not affect at least two, if not all three, of these factors as part of their natural evocation. Thus, the multifaceted empirical strategies of social psychologists help to illuminate behavioral phenomena *in toto*. Both economics and social psychology could benefit by attempting to consistently and conscientiously measure all three of these outcomes in their experiments.

Multiple Methods

Imagine that you are a football coach who wants to determine the optimal amount of time to practice between games. In professional football in the United States, your regular season is 16 games in 17 weeks (once during the season you have two weeks between games). You know that more practice will help your team members learn the plays better and coordinate with each other more effectively, but it will also increase their risk of injury, it might sap their strength, and too much practice might have a dulling effect on their maximum motivation and creativity. (As classical musicians put it, when they talk about how much to rehearse, "you must leave some room for the tie and tails," that is, some room for spontaneity.)

An obvious solution to this problem is to do an experiment by practicing more during some weeks and less in others. (This is clearly a risky strategy given the competitive nature of professional sports and the perilous nature of your own employment.) The bigger problem methodologically is that you can't control many of the other factors that might influence the effectiveness of your practice: You don't play the same team each week and even though you play a small set of teams twice during the year, the composition and possibly the motivation of those teams

changes due to injuries and other contextual effects. Not only that, your own team changes from week to week as players get hurt and you demote and promote others depending on their effectiveness.

One counter-normative way to solve the problem would be to have several teams run similar experiments and compare all of the results. This would help to mitigate the effects of changes in teams and contexts and might focus the overall effects more directly on each team's practice times.

This example, simple as it is, suggests two conclusions for social psychologists and experimental economists: Controlling extraneous factors, as we all know, is critical, and a single experiment may not be enough to provide firm conclusions. A series of conceptual replications that produce consistent results can increase confidence in the internal validity of the findings.

Our methods could go even further if we also want to establish the external validity of our findings. At a dinner in his honor in 2006, Gary Becker noted that, although our methods and techniques may well change over time, the goal of economics will remain the same, that is, "It is judged ultimately by how well it helps us understand the world, and how well we can help improve it." von Neumann (1944, p. 2) also noted the value of different methods: "The differences in scientific questions make it necessary to employ varying methods. . . ."

In this chapter, I would like to reiterate some of what I learned (but have only rarely practiced) in graduate school, that is, that many questions are best addressed not only by conducting multiple experiments but also by using more than a single method. Thus, the utilization of simulations, field surveys, or interviews, in addition to experiments, can strengthen our work's empirical foundations when we test theoretical predictions and hypotheses.

My central, primary source in my methodological training in graduate school was Philip Runkel and Joseph McGrath's (1972) amazing book, *Research on Human Behavior: A Systematic Guide to Method*. They identified eight research strategies that they arranged in a circle and categorized on two dimensions, obtrusive–unobtrusive research operations and universal–particular behavior systems (see Figure 10.1). They also classified the eight strategies in terms of the setting in which behavior was observed (if behavior was observed at all) and in terms of a researcher's concerns for generality, precision, or the context itself. Table 10.1 provides definitions for all of the eight strategies.

Runkel and McGrath characterized laboratory experiments as involving obtrusive research operations in contrived and creative settings, designed to reflect, at least slightly, universal behavior systems, with tremendous concern for precise behavioral measures. Although they spend more time on experiments than any of the other strategies, Runkel and McGrath seem completely unbiased toward any of the eight strategies, as they note, "We cannot emphasize too strongly our belief that none of these strategies has any natural or scientific claim to greater respect from researchers than any other" (p. 89). Instead, they view the different approaches as complementary and in service of each other: whatever strengths one strategy might

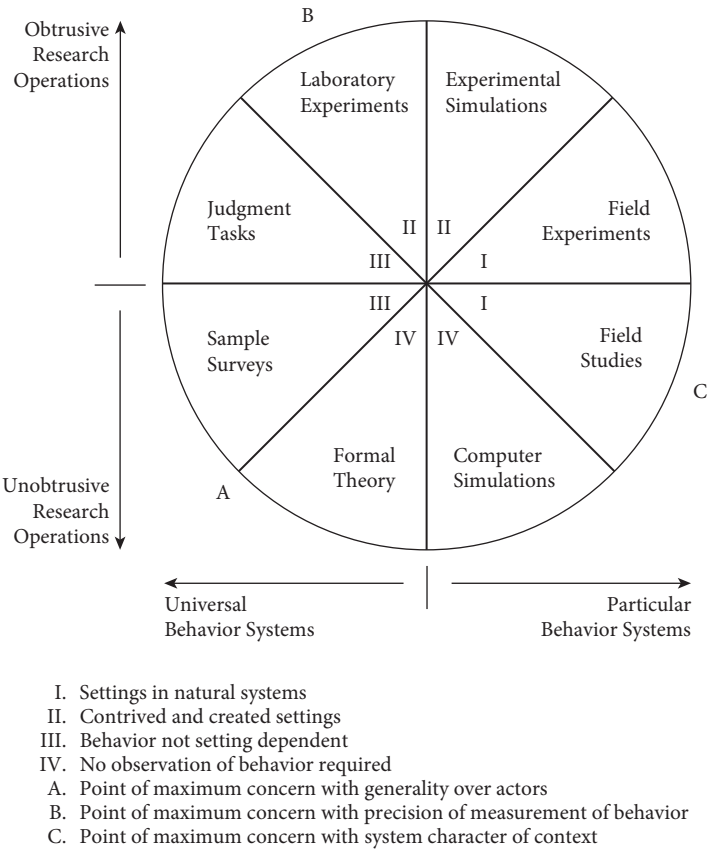


Figure 10.1. A framework for comparing some major research strategies. Figure originally appeared as Figure 4.1 in Runkel and McGrath (1972, p. 85). Reprinted with the permission of the publisher.

have, another might compensate for its weaknesses, and each strategy has both strengths and weaknesses.

They define a laboratory experiment as involving "the study of one or more behavioral processes under conditions highly controlled by the investigator" (p. 103). The intent is "to exemplify generically or prototypically a cluster of processes, quite apart from the settings or systems in which those *processes* are found naturally" (p. 104, emphasis in the original). The goal for a lab researcher is to learn "a great deal about a very narrow scope" (p. 107). They note that experiments are designed to maximize internal validity and, as a result, may sacrifice external validity, two concepts of central importance in another classic methods primer, Donald Campbell and Julian Stanley's (1963) *Experimental and Quasi-experimental Designs for Research*. The dangers in running experiments results when "the laboratory becomes a very special social system in itself with an explicit or implicit agreement between actor and researcher that neither is going to behavior 'naturally.'"

Table 10.1. Research Strategies Described in Runkel and McGrath (1972)

Strategy	Definition
Field studies	Systematic observations of behaviors within naturally occurring behavioral systems with as little disturbance of those systems as possible.
Field experiments	Systematic observations of behaviors within naturally occurring behavioral systems in which the experimenter deliberately manipulates one or more variables.
Experimental simulations	Observations of behaviors within deliberately constructed settings that simulate some class of naturally occurring settings as nearly as possible.
Laboratory experiments	The study of one or more behavioral processes under conditions highly controlled by the investigator.
Judgment tasks	Studies in which the investigator carefully establishes control conditions within which actors whom the investigator believes to be relatively uninfluenced by their origins or backgrounds will render the judgments the investigator seeks.
Sample surveys	Studies in which the investigator seeks a specified partition or random sampling of actors to render judgments the investigator believes to be relatively uninfluenced by context.
Formal theory	The construction of an abstract, logical model of a behavior system—usually a generic class of behavior systems—with logical manipulations (e.g., mathematics) to adduce new insights.
Computer simulations	Observations of behaviors within deliberately constructed settings that simulate some class of naturally occurring settings as nearly as possible in an artificially complete, closed system.

The result is that "the behavioral scientist can, literally, produce behavior that is indigenous only to the laboratory" (p. 106).

The major difficulty with any of the eight research strategies is that "realism, precision, and generality . . . *cannot* be maximized at the same time" (p. 115, emphasis in original). The advantage of running experiments is their precision, which is particularly important in the testing of well-specified economic theories. At the same time, however, they often suffer from a lack of realism and a lack of generality. When theory testing is the sole purpose of the empirics, then experiments are a perfect tool. It is only when we want them to be more than theoretical testers that they need bolstering.

Certainly in business schools, we are often asked to step out of our offices and discuss the implications of our research. Economists have also been occasionally asked about the relevance of their research for public policy. For us to say more than "this theory was supported in these conditions" with confidence, it may be necessary to utilize research strategies that will complement our experiments. Runkel and McGrath's framework suggests that sample surveys, field studies, and

field experiments might all help to increase both the realism and generality of our findings—if we replicate our findings. Not replicating opens a whole host of methodological questions and quandaries—and the likelihood of nonreplication is potentially quite high, especially, as noted earlier, if participants are transforming our instructions (Kelley and Thibaut, 1978). Nevertheless, to achieve any strong basis for making claims of external validity, methods other than experiments may be necessary.

Another approach implied by Runkel and McGrath's circumplex model might be to combine the findings of laboratory experiments with formal theory and field experiments—a direction that economics seems to be pursuing to great advantage (e.g., Harrison and List, 2004; List and Reiley, 2002). The sometimes limited opportunities for field experiments reduce the applicability of this approach and, ironically, suggest methods like field studies and sample surveys that may be somewhat unpalatable because of their potential lack of precision. When the questions that we address demand that our findings have more than internal validity, however, we might need to consider the contributions that less precise but more general methods can provide. From a social psychological rather than an economic point of view, the call for more formal theory as well as an appreciation of alternative empirical methods should be one that is embraced rather than resisted, as seems to be the current reaction of the field.

Conclusions

The bottom line here focuses on several issues. First, when we are investigating human behavior, cognition, and emotion, we are dealing with complex phenomena. Second, economists and social psychologists have a plethora of common interests and a common affinity for experiments. Third, we are limited when we use a single experiment to address a research question; No single experiment can be sufficiently definitive to confidently answer the kinds of complicated questions that we often ask. Fourth, we have many audiences who would like to hear the implications of our work; to be responsible when we share it with them, we must ascertain the reliability and both the internal and external validity of our findings. Fifth, beneath our research cloaks, we are also human beings, with emotions that encourage us to believe our own theories even when our experiments have limitations and alternative explanations. Sixth, we tend not to be attracted to or trained in multiple methods, much less multiple disciplines; this limits our ability to conceptualize important interdisciplinary questions and to investigate key phenomena in multiple ways. Seventh, we have called for some time for these two connected fields to do more together, but progress has been infuriatingly slow.

The ultimate conclusion: There is work to be done, in all of our laboratories. We would do well to collaborate with each other more, to investigate questions that are attractive to both fields, to learn a wider array of methods and research

strategies, to get out of our labs and get our hands messy in the field, and to be provocative in our research questions and cautious with our conclusions.

REFERENCES

Allport, A. 1985. The Historical Background of Social Psychology. In *Handbook of Social Psychology*, eds. G. Lindzey and E. Aronson, Volume 1, 3rd edition, New York: Random House, pp. 1–46.

Berg, J., J. Dickhaut, and K. McCabe. 1995. Trust, Reciprocity and Social History. *Games and Economic Behavior* **10**:122–42.

Camerer, C. and R. Thaler. 1995. Ultimatums, Dictators, and Manners. *Journal of Economic Perspectives* **9**:209–219.

Campbell, D. and J. Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Boston: Houghton Mifflin.

Croson, R. 2006. Contrasting Methods and Comparative Findings in Psychology and Economics. In *Social Psychology and Economics*, eds. D. De Cremer, M. Zeelenberg, and J. K. Murnighan. Mahwah, NJ: Lawrence Erlbaum, Inc., pp. 213–234.

Gilovich, T., R. Valone, and A. Tversky. 1985. The Hot Hand in Basketball: On the Misrepresentation of Random Sequences. *Cognitive Psychology* **17**:295–314.

Harrison, G. W. and J. A. List. 2004. Field Experiments. *Journal of Economic Literature* **42**:1009–55.

James, W. 1890. *The Principles of Psychology*. New York: H. Holt.

Jamison, J., D. Karlan, and L. Schechter. 2007. To Deceive or Not to Deceive: The Effect of Deception on Behavior in Future Laboratory Experiments. Unpublished Manuscript, University of California, Berkeley.

Kahneman, D. and A. Tversky. 1971. Belief in the Law of Small Numbers. *Psychological Bulletin* **2**:105–110.

Kelley, H. H. and J. Thibaut. 1978. *Interpersonal Relations: A Theory of Interdependence*. New York: Wiley.

Ku, G., A. Galinsky, and J. K. Murnighan. 2006. Starting Low but Ending High: A Reversal of the Anchoring Effect in Auctions. *Journal of Personality and Social Psychology*, **90**:975–986.

List, J. A. and D. H. Reiley. 2002. Bidding Behavior and Decision Costs in Field Experiments. *Economic Inquiry* **40**:611–619.

Lount, R. B. and J. K. Murnighan. 2007. Can Positive Mood Impact Trust Development? Paper presented at the EIASM Conference on Trust in Amsterdam, The Netherlands.

Murnighan, J. K. and T. Ross. 1999. On the Collaborative Potential of Psychology and Economics. (The introduction to a special issue). *Journal of Economic Behavior and Organization* **39**:1–10.

Platt, J. 1964. Strong Inference. *Science* **146**:347–353.

Rapoport, A. and A. M. Chammah. 1965. *Prisoner's Dilemma*. Ann Arbor: The University of Michigan Press.

Runkel, P. and J. McGrath. 1992. Research on Human Behavior: A Systematic Guide to Method. New York: Holt, Rinehart, & Winston.

Samuelson, P. 1948. *Economics*. New York: McGraw-Hill.

Skinner, B. F. 1953. *Science and Human Behavior*. Upper Saddle River, NJ: Pearson Education.

Staw, B. 1976. Knee-Deep in the Big Muddy: A Study of Escalating Commitment to a Chosen Course of Action. *Organizational Behavior and Human Performance* **16**:27–44.

Tversky A. and T. Gilovich. 1989. The Cold Facts About the Hot Hand in Basketball. *Chance* **2**(1):16–21.

Tversky, A. and D. Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science*, **185**:1124–1131.

von Neumann, J. and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Zhong, C., J. Loewenstein, and J. K. Murnighan. 2007. Speaking the Same Language: The Cooperative Effects of Labeling in the Prisoners' Dilemma. *Journal of Conflict Resolution* **51**:431–456.

CHAPTER 11

PSYCHOLOGY AND ECONOMICS: AREAS OF CONVERGENCE AND DIFFERENCE

TOM R. TYLER
AND DAVID M. AMODIO

Introduction

The title of this session is "psychology and economics: A clash of methods?" We see both convergences and differences in the approaches taken to methodology within our two disciplines. We are uncertain whether they amount to a clash of methods, but we agree that they reflect different goals in designing research. In this chapter, we highlight what we view as three of these differences: the focus of study, what is appropriate research design, and issues involved in the study of economics in the context of the brain. At its core, the "clash" of methods appears to concern issues of experimental validity that arise when attempting to infer underlying (i.e., "latent") psychological variables from observable (i.e., "manifest") responses, such as behavior or physiology. We note that while economists' interests have moved toward a greater focus on underlying psychological constructs that are not directly observable, their experimental practices require updating to deal with the new challenges of psychological investigation.

What Should Be the Focus of Study—Psychological Constructs or Observable Behavior?

Before considering a clash of methods, it may help to consider the differences in the overarching goals of research in behavioral economics and psychology. Traditionally, behavioral economics seeks to develop models to predict decisions about money and monetary exchanges. For example, experimental games examine when people will trust others as is reflected by taking risks based upon anticipated cooperation from those others. By comparison, psychology seeks to understand the emotional and cognitive mechanisms that underlie complex behaviors. For example, during an intergroup interaction, how might one's implicit prejudice toward a person interact with concerns about social desirability to predict the friendliness of a response? Hence, the traditional focus of each of these two fields is very different: Economics focuses on decision outcomes, and psychology focuses on mental processes.

The methods in these two fields evolved to support their respective traditional research goals. However, as we will discuss, the goals of economics have moved considerably toward those of psychology in recent years, particularly given the emergence of behavioral economics and neuroeconomics; as a result, it may be necessary to reassess these methods.

When it comes to experimental approaches, economists have traditionally focused on behavior, using observed behavior to "reveal" people's preferences. While "preferences" are mental states, they are not typically measured directly and are instead inferred from behavior. Of course, like all generalizations, there are exceptions, and economists increasingly do directly measure some mental states—for example, beliefs and expectations (Fehr, 2007).

Historically, the study of observable behavior has also been central to psychological research. For example, during the classic era of learning theory, such famous psychologists as Skinner placed the same strong emphasis on behavior widely seen in economics. Of course, Skinner primarily studied animals, so his direct access to mental processes was, by definition, limited. However, when he did venture into the human realm, Skinner continued to focus upon observed behavior and, as a matter of principle, did not make hypotheses about mental processes. Since the fall of Behaviorism (Skinner's theoretical framework), other psychologists, while continuing to study behavior, have also been open to trying to directly measure people's mental states (i.e., their attitudes, values, and feelings)—the psychological processes behind the behavior.

Since the cognitive revolution in the 1960s there has been a broad consensus within psychology that it is important to study mental processes. Psychologists have come to believe that people employ complex representations of the environment, that they use processes of construction and reasoning to actively

create mental representations, and that emotions influence their decisions. From the psychological perspective, these can be best understood by direct measurement. One example of such measurement is by self-report. For example, a research subject might report in a questionnaire the extent to which different thoughts or emotions are active during the process of making a particular decision. Another approach involves measuring some aspect of physiological or neurological activity, from which the researcher might infer the activation of an emotional or motivational state during a decision-making process.

Sometimes this tendency goes too far, such as when psychologists focus exclusively upon internal mental states, ignoring behavior. To the degree that behavioral economics has had the effect of refocusing psychologists' attention upon behavior, it serves as a useful corrective to such psychology. Those psychologists who study mental processes often acknowledge the problems involved in considering subjective judgments, and they frequently point out their concern that mental processes may not map cleanly onto behaviors. Since as early as the 1930s, psychologists have been concerned with the disconnect between measures of psychological processes and actual behaviors. For example, a classic paper from social psychology demonstrated that people's attitudes, no matter how strongly they were stated, were not always found to correspond with their actions in social settings (LaPiere, 1934).

The primary goal of psychology is to understand mechanisms of the human mind and its ability to interface adaptively with the (social) world. Although only behavior and physiology are objectively observable, it is assumed that observable behaviors are driven by internal processes. It is recognized in psychology that it is difficult to understand patterns of behavior without an effort to understand the rich and complex mental processes that characterize the human experience. As a consequence, psychologists, while not averse to studying behavior, are also strongly committed to measuring people's thoughts, attitudes, and feelings. This does not in any way mean that psychologists are unaware of the subjectivity involved in measuring unobservable mental constructs. Economists often choose to avoid the potential ambiguities introduced by more subjective measures by dismissing the measurement of mental states altogether.

Psychologists do not dispute that these problems exist—indeed, many psychologists have been active in addressing these issues, such as in the area of psychometrics and psychological methods. Psychologists are interested in identifying the causes of behavior, rather than merely describing patterns of observed behavior. They are committed to searching in an area in which they think important information can be found even though it is an inherently more error-filled process than using a more objective measurement tool but at the expense of missing the critical underlying mechanism. Given this state of affairs, a great deal of psychological research is directed at understanding what type of inferential errors occur within different types of research and, where possible, how they can be corrected. Psychologists see these issues as a necessary part of addressing questions about the mechanisms underlying human behavior.

While economists typically take a strong pro-behavior stance in theory, in our experience that stance is not held in practice. Economists have a widespread tendency to explain their findings by invoking psychological states, without providing any evidence to support their "psychologizing." Let us give two examples. First, trust games. To psychologists, "trust" is a state of mind. I do or do not trust you. That state of mind shapes my behavior. But, does that mean that we can observe patterns of behavior and label them as reflections of trust? Similarly, justice is something that exists within the mind of an individual as a subjective judgment. When economists observe a pattern of behavior and label it "trust" or "inequity aversion," they are inferring that the behavior they see is motivated by people's concerns about justice. In other words, they are using a term that implies a subjective state to account for an observed phenomenon.

But what evidence shows that these labels are correct? Are economists using these descriptions of mental states as mere labels of objectively observed behavior? Or do they think that these titles reflect explanations for why the observed behavioral effects occur? If economists do not mean to be making inferences to mental states that they are suggesting cause observed behavior, it seems strange for them to so frequently rely on mental state labels as descriptors. If economists cannot describe behavioral patterns without inferring mental states, then perhaps this suggests that they should take the study of mental states seriously.

The problem is that statements that infer but do not measure a mental state cannot be falsified. It is difficult to determine whether "trust" or "justice" or "mood" is the actual mechanism which leads to a behavior without measuring that psychological characteristic directly or finding some indirect evidence that these states, rather than some other, are responsible for the observed behavior. Psychologists who directly measure such mental states can test whether or not they actually mediate the relationship between stimulus and response. If such states are not measured, it is more difficult to infer the mental processes that are occurring.

This is not to say that it is not sometimes possible to infer mental states without directly measuring them. For example, we can and do manipulate the situation to create circumstances under which we think that justice motivations should be activated, or emotions aroused, and then look for a predicted behavior under those, but not other, predicted conditions. In other words, it is possible to test psychological explanations without measuring mental states through the use of careful experimental designs. But, even here psychologists emphasize the use of "manipulation checks" to see if the conditions manipulated actually influence the hypothesized psychological state. This does, however, seem like at best a partial effort toward inferring a particular mental states underlies a behavior.

In the end, we think that behavioral economists would benefit it they would think first about how to directly measure mental states and second about theoretical issues relating to the role of mental states. Or, if they do not wish to do so, it would be helpful if they would be sensitive to the problems created by using psychological terms, such as altruism, reciprocal kindness, trust, and inequity aversion if there is no evidence for the use of that term in particular.

The Participant as a Self-Aware Entity

Irrespective of whether we study behavior or mental states, psychologists are traditionally very concerned about the self-awareness of their participants. This self-awareness leads participants to know that they are being studied and to say and do the things that create an impression of themselves that they want to project to the social world. Such self-awareness is an issue because it is a plausible rival hypothesis to the argument that the treatment caused whatever experimental effects are observed. So, for example, when a subject sees smoke coming under a door, he might become afraid; this fear drives his behavior. But, it might also be the case that the subject thinks, "What is the experimenter trying to accomplish in this study by blowing smoke under the door and how can I look good in her eyes?" If, in fact, the subject is acting based upon his inferences about the goals of the study, and not a true reaction to the stimulus smoke, then the experiment makes incorrect inferences about psychological processes.

Psychologists refer to this phenomenon as *participant bias* or *demand characteristics*, and it constitutes a major threat to the validity of the inferences that one can draw from an experiment. If we ask people how often they break the law or engage in deviant sex, we see self-reported answers that we suspect reflect people's desire to look like a good person. For example, in studies of obedience to the law, people report following the law more frequently than studies of the rate of crime suggest is possible (see Tyler (2006b)). Similarly, when people think they are being observed, they are more likely to help someone in need, less likely to steal something, and so on. Furthermore, people are likely to behave as they think the experimenter *wants* them to behave, even when there is no issue of their self-presentation.

Like the uncertainty principle in physics, the act of observation changes the phenomenon that is being observed. Psychologists have adopted a number of approaches to trying to deal with this basic social reality, and some of those have been criticized by economists. The deception that is sometimes used in psychology experiments seems to be a particular target. We will address other approaches later in this chapter. But, first we want to ask a basic question of behavioral economists. Based upon the design of experiments in behavioral economics, it would seem that economists do not feel that this is an issue that needs to be addressed. For example, in a multi-trial dictator game, psychologists argue that we need to be concerned that people might be motivated by self-presentation and might, as a result, behave in ways that do not reflect how they would behave outside of the laboratory? To the extent that participants are concerned with self-presentation, as extensive research has shown, psychologists suggest that we must ask how researchers should deal with those effects.

One approach is to do studies in the real world, like the anthropologists who have taken the dictator game to other societies (Gintis et al., 2005; Henrich

et al., 2004). In such settings, are we more confident that people are acting as they would in their natural lives? Is playing a game with an anthropologist, whom you might want to impress with the virtue of your society, really the way to understand how people will act when they go to the market later that week? In other words, the problem of self-presentation is central to social acts, irrespective of whether they go on in a laboratory or in the field. As Glenn Harrison observed in his presentation at the conference from which this book emerged: "How can we be sure [that] subjects are playing the game that we think they are playing?"

Our general point is that one must consider the broad variety of goals and motivations that participants have that operate when they are behaving outside of the narrow context of the experimental situation; and, in particular, we must ask if those goals and motivations are the same as the ones that people have when participating in research. Of course, one difference between traditional economic research and social psychological research is that economists focused on money and material goods, whereas psychologists studied issues such as love, altruism, and aggression—issues in which people can easily be seen to be interested in the impression they make upon others. However, more recently, economists seem to have expanded the scope of their concerns to include more social motivations, and this suggests that they need to be more sensitive to the issue of "self" and the self-concerns that people bring into experiments. They have also made greater use of studies of undergraduates, which is an issue of long-term concern to psychologists (Sears, 1986).

In our view, economic decision-making games presented as objective indicators of choice are often laden with interpersonal motivations and emotions. For example, the ultimatum game is not a game simply about money. It raises issues of justice, which are linked to self-image and to self-esteem. A core principle of many social groups is that resources should be distributed fairly. While earning resources may entitle someone to more than an equal share, the ultimatum game typically presents people with a case of unearned resources. This creates strong pressure to distribute resources equally, and not doing so reflects poorly on a person in their own eyes and those of others. Hence, self-presentation issues are central to people's behavior. In general, psychologists find social motivations are central to a wide variety of types of behavior in social settings (Tyler, 2006a; Tyler and De Cremer, 2006a,b).

How can we deal with such self-presentational motivations in our experiments? In technical terms, how can we mitigate participant bias and thus preserve the internal validity of our experiments? One way is to disguise that people are being studied. This is called unobtrusive measurement (Webb et al., 1966). It includes indirect measurement, behavioral observation, and physiological assessment, all mechanisms designed to minimize the ability of people to craft their behavior to create a desired impression. The second approach is to disguise what is being studied, for example, in a deception experiment (Aronson et al., 1990). In our

experience, there is no subject that provokes so much emotion among economists as the use of deception in an experiment.

Of course, psychologists do not do deception because they are uncivilized and amoral people. They do it in part to address the critical issue of internal validity—without which conclusions from an experiment cannot be drawn (Campbell and Stanley, 1963). That is, they fear that if the subjects know what the study is really about, they will decide how they want to present themselves and will give verbal reports and actual behaviors to support that self-image. The concern is that when subjects are not in a study and not motivated by such self-presentation concerns, they will act differently. So, for example, we may see very caring and helpful participants in a lab setting, but the same people will not give a dime to beggars on the street (unless they think they are still in the study) (see Batson (1991)).

Of course, psychologists also have other reasons for conducting experiments involving deception. For example, they may want to create particular feelings, thoughts, or goals that people experience in real-world settings in a way that can be studied in the laboratory. This refers to ecological validity—the extent to which the manipulations of social factors in an experiment seem natural and uncontrived to the participant. An experiment with good ecological validity will produce results that are more predictive of real-life behavior outside the lab.

Do economists have a way to address these problems? Or are they simply ignoring them? If economists are ignoring them, then we need to ask if that is a problem. One perspective on this question is that behavior is behavior, and the motivations underlying that behavior are not a key issue. From a different perspective: If people's behavior under a given set of conditions in a laboratory study is different from their behavior under similar conditions in which they are not being studied, then there is a problem. The subjects in behavioral economics studies know that they are in a lab and in a study. Does that change their behavior? Has there been any research on this?

In a broader sense, there is a large area of psychological research design that grapples with the question of how to design studies to minimize these problems (Aronson et al., 1990). There are basically two ways to deal with this issue. One is deception. Mislead subjects about the true purpose of the study so that they have trouble behaving in ways that might undermine the internal validity of the study. The other is indirect measurement. Measure things in ways that make it hard for the participant to control their responses. For example, we might use a projective test. This concern leads to unobtrusive measurement. People do not know that they are in a study or, if they do, they do not know what is being measured. All of these ideas are in the service of establishing internal validity.

Setting aside for a moment the many critical functions of deception in psychological experiments, we acknowledge that there are also practical concerns that the use of deception in one experiment will "spoil the subject pool" for subsequent experiment. That is, if a subject learns that deception is commonly used in

experiments, then he will no longer act naturally in future studies. This is an issue that psychologists must deal with as well. Certainly, psychology research subjects are usually aware that deception is often used in psychology experiments. As a result, they may be vigilant for signs of deception, trying to decipher the researcher's "true" intent.

Psychologists are well aware of this issue. But given the very basic and critical need for solid internal and ecological validity in their studies, most psychologists have decided that deception is necessary. Therefore, over the past half-century, the field has developed several clever and effective methods for overcoming the potential problems of the vigilant and skeptical participant. For example, it is common to mask the true intent of an experiment with a very plausible and compelling alternative purpose, often with the use of a clever cover story. In addition, it is common practice in social psychological studies to use a "funneled' debriefing format, in which a subject is carefully and systematically probed for suspicion about the cover story and the true hypothesis being tested in the study. Speaking from our personal experience, it is most often the case that undergraduate participants are neither vigilant of nor privy to the use of deception in our experiments. Thus, there are well-established methods for address the concern of "spoiling the pool"; and at a practical level, the concerns expressed by economists about "tainting" people for future studies appear to be exaggerated.

As economics moves further into the study of social motivations, it is inevitable that it will have to address the broader issues we are raising. While it may seem unusual to an economist to argue that people care more about their identity and self image than they do about their material rewards and costs, such arguments are not unusual in psychology (Tyler, 2011), and they are becoming more commonplace in economics (Akerloff, 2006). Indeed, the surge in interest in behavioral economic decision-making games, such as the Ultimatum Game and the Dictator Game, are founded on the observation that people appear to care more about social norms of fairness and self-presentation than they do about money—the concern about social factors over money is precisely what makes these games so interesting to behavioral economists. From our perspective, it appears that economists are becoming social psychologists in their questions, but still need to consider how to develop the statistical and methodological tools needed to study such questions in a scientifically valid way.

Economics and the Brain

Neuroeconomics is a rapidly emerging area of economic studies (Glimcher et al., 2008). The allure of neuroeconomics is that neuroimaging methods, such as functional magnetic resonance imaging (fMRI), may provide a direct window into people's thoughts and emotions. There are, of course, parallel literatures in psychology in cognitive neuroscience and social neuroscience. The fascination with

neural processes is shared by our two fields. Interestingly, when it comes to neuroimaging approaches, the "clash" in methods is between the new areas of cognitive neuroscience and the traditional fields of social and cognitive psychology.

Neuroimaging, and psychological methods more broadly, are appealing for several reasons. A major reason, most relevant to the present discussion, is that neuroimaging methods may provide a more direct assessment of internal states related to decision making and choice. The hope is that a direct neural probe would allow one to circumvent the various social and self-presentation biases described above. Although this technique may seem new, psychologists have gone down this road before, most notably with the use of the polygraph in lie detection. A century's worth of research has investigated the extent to which various physiological indicators, such as the galvanic skin response (i.e., palm sweating), heart rate, or nonverbal behaviors (e.g., eye blinks, gaze aversion, etc.), can discern truth from a lie.

Inferring "True" Feelings from Physiology: The Case of the Polygraph

The idea that we can bypass people's self-reports, which are generally under the subject's control, and get directly at their true feelings, or that we can observe aspects of behavior that the person cannot control or is not aware of, again to get at their true feelings, solves so many scientific and societal problems that it has broad intuitive appeal. That appeal is well illustrated in the area of crime and law. In the legal arena, people are often highly motivated to lie and obscure. Yet theories of justice are based upon knowing the truth—that is, who is guilty and what did they do? Horrific stories emerge from the middle ages of hot swords placed upon people's tongues or of people being ordered to swallow dry bread to determine whether a person was lying. These crude lie detectors are based upon the intuition that people who are lying have a dry mouth. Although these methods seems silly to us now, in their conceptualization they are identical to the modern lie detector.

We dwell on the lie detector because it represents the same problems that we see with studies using neuroimaging and other physiological methods. In each case, the issue concerns what the physiological indicator is measuring—that is, its *construct validity*. Lie detectors do not measure lying; they measure autonomic arousal, or, more precisely, changes in heart rate and in palm sweating (in the case of galvanic skin response). If an innocent person is anxious because they fear false accusation, and this produces a physiological response like that of a guilty person who is anxious because they fear true detection, then the innocent man will be charged as guilty. The point is a key one: Physiological measures must be labeled as to what they indicate to be useful. After a century of research on lie detection, the recent Intelligence Science Board review of mechanical methods of detecting deception

concluded: "None of these mechanical devices has been scientifically shown to be capable of accurately and reliably detecting deception" (p. 83). Why not? We cannot be confident in interpreting the meaning of the measured physiological responses. The measures lack construct validity.

The polygraph was also heralded as measuring responses that were beyond one's deliberative control—a direct pipeline to one's true knowledge and intent. But this is not the case. People can be trained to "beat a polygraph" by altering their physiology in a variety of ways. While secret training of spies and special forces soldiers in counter tactics to defeat the polygraph is the stuff of films, readily available manuals offer the everyday criminal a set of practical techniques for this task such as randomly curling your toes and sticking a tack into your hand. Similarly, every time a new study of how to detect behavioral signs of deception is published, a new version of counter guidelines becomes available concerning tactics to undermine it ("look people in the eye"; "speak more slowly"). These behaviors may or may not be under conscious control and there is an ongoing discussion about what people can and cannot bring under their control. Nevertheless, it is clear that physiological and neural responses can reflect self-presentation efforts.

What does this have to do with neuroeconomics? Economists want to draw inferences about psychological constructs such as justice and trust without acknowledging that such complex psychological processes cannot be directly inferred from a simple pattern of brain activity or physiological responses. This is not the fault of economists—it is a common mistake that has repeated itself throughout the history of psychology in various forms, with the most recent case being human cognitive neuroimaging—but we want to alert economists to the problem.

The Supreme Court recently provided a practical example of this problem when it decided that juvenile offenders could not be executed, citing evidence that the regions of the brain responsible for decision making and impulse control are not as well developed in adolescents as in adults. This case illustrates the allure of relying upon science to make the decision, and it certainly shows its likely legitimating power. Several recent reports have shown that neuroscientific evidence is erroneously treated as hard evidence (McCabe and Castel, 2008; Weisberg et al., 2008). As observed by Aronson (2007, p. 133): "If you can show that this part of the brain is active, and this part of the brain is not, it is hard science." But, as with the lie detector, the problem lies in interpreting this evidence. In actuality, the results of brain imaging studies are much more ambiguous and unclear than is suggested by their description as "hard science." For example, in this particular case, there is no direct evidence linking any particular measure of brain functioning to how decisions to commit a crime are made. The Supreme Court decision may have made political sense, given the mystique of neural science, but was it good science?

Of course, neuroimaging methods such as fMRI are now being used in the case of detecting deception, supplanting the lowly lie detector and its many problems. But, a recent review of brain scanning methods by the Intelligence Science Board

suggests that "it is unlikely that a behavior as complex as deception can currently be diagnosed with any of the existing brain imaging techniques" (p. 80). Their review considers EEG, MEG, PET, fMRI, fNIRS, and TMS techniques and suggests that none has been found to be an accurate and reliable detector of deception. Indeed, the same problems of construct validity that plagued the polygraph also plague fMRI. There is simply not at this time a reliable way to map high-level psychological constructs, such as deception, onto specific patterns of physiology or neural activity.

Neuroimaging in Behavioral Economics

Without a doubt, the effort to integrate neuroscience into behavioral economics holds much promise. But there are still many obstacles to overcome before neuroimaging studies will yield substantive advances in the sciences (Amodio, 2010). At the core of these obstacles is the issue of construct validity—the same problem that has plagued efforts in physiological lie detection, as described above, as well as previous efforts in psychology to develop measures of underlying personality styles and, before that, unconscious motives (Cronbach and Meehl, 1955). In each case, the measurement tool did not always provide a valid index of the psychological process of interest. The biggest problem with neuroimaging approaches is that researchers too often assume a one-to-one mapping between psychological constructs and physiology, when in fact the organization of the brain may be very different than the way scientists believe human thought and emotion is organized (Cacioppo et al., 2003).

As noted above, the psychological construct of "trust" is of great interest to economists. Although trust might be measured most directly through a self-report questionnaire (e.g., "How much do you trust your partner?"), a researcher might worry that self-reports may be biased by self-presentational concerns and thus seek an alternative measure that does not rely on self-report. Can trust be measured using fMRI? This would require that a particular neural structure, or combination of structures, is dedicated to the rather complex and high-level psychological construct of trust—an assumption that seems very unlikely.

Studies that have examined the neural correlates of trust have suggested that the amygdala provides a marker of distrust (Adolphs et al., 1998; Engell et al., 2007) whereas regions of the basal ganglia (e.g., the striatum), often linked to reward, provide a marker of trust (Delgado et al., 2005; Baumgartner et al., 2008). However, these neural structures are extremely complex, with multifaceted functions that likely go far beyond the very specific emotional appraisals that are inferred by researchers during a subjects' economic game play. The amygdala is not simply a fear module—it represents a complex network of subnuclei that are responsive

to both threatening and rewarding stimuli, and it appears to function to orient attention for goal-directed behaviors and to modulate activity of the autonomic nervous system to prepare for fight-or-flight responses. Although the amygdala may be associated with the broad set of psychological, physiological, and behavioral processes related to a trust-based decision, it would be a reversal in inferential logic to claim that amygdala activation reveals a neural substrate of distrust (Poldrack, 2006). Similarly, the striatum has been implicated in a wide range of processes that are not clearly related to trust, such as implicit procedural memory, motor coordination, and response anticipation (e.g., Yin and Knowlton, 2006). Given these complexities, the inference that activity in the amygdala or striatum reveals trust is very speculative.

A more basic construct of interest to economics is the notion of value. Value can be defined objectively in terms of amount of money (e.g., the value of a $5 bill). Economists are particularly interested in how humans understand and respond to value. In neuroeconomics, the notion that there is a special place in the brain that computes value has become very popular. The hope is that if we can measure the activity in this part of the brain, then we can understand most economic decisions made by humans. To identify the "value computer" in the brain, researchers have scanned participants while considering the value of different stimuli in experimental tasks, ranging from taste tests (Plassmann et al., 2008) to rather complex decision-making games (Montague and Berns, 2002). Consider, for example, in a highly cited paper that purported to reveal the neural structures that compute expected value (Knutson et al., 2005). In it, participants' brains were scanned while they viewed a series of different shapes, followed always by a white square, a button press, and then feedback on whether money was lost or gained on that particular trial. Gains and losses ranged from ± $1 to $5. The authors noted which brain regions were more strongly activated on trials where the participant prepared to respond to a shape that promised to yield a larger reward (versus larger loss). Whatever regions showed activity associated with this statistical contrast were deemed to be the neural substrates of value. Other researchers have used similar methods. To us, it seems difficult to look at brain activity from this type of experimental task—which involves a wide range of psychological processes such as learning and memory, attention, motor coordination, emotion regulation, and so forth—and infer that whatever lights up represents the single process of expected value computation.

In contrast to the economist's view that decisions involve a single computation of value, decades of research in psychology have classified unique subcomponents of the decision-making process into those involving predecisional weighing of options, response planning, response implementation, and postdecisional assessment, among others (Gollwitzer, 1990). Progression through these different phases of decision and action can occur quickly, and it is notable that the slow timescale of hemodynamic measurement in fMRI likely blurs these different stages of the decision-making process.

Furthermore, economic decision-making experiments have not yet been designed properly to disentangle the different components of decision and action (at least according to psychological models of these processes). As a result, it is virtually impossible to know what aspect of the decision-making process a brain activation in such studies might represent. For example, is an observation of striatum activity related to the weighing of options? The resolution of a decision? The planning of a movement to implement the decision (e.g., to make a response on a button box)? The implementation of the movement? The activation of a previously learned procedural script? The satisfaction of making a decision? The anticipation of an upcoming reward? Preparation to receive a reward? The list of plausible alternatives goes on, and a particular brain activation (e.g., in the striatum) could reflect any of these processes. If one's theory assumes that behavior on a trial involves only one specific process, it may lead to an incomplete and potentially incorrect interpretation of brain activity. Although we note this problem in the context of neuroeconomics, it is a problem that applies to cognitive neuroscience more generally. From a psychological point of view, the ascription of complex psychological processes such as trust or value to a particular brain structure is virtually impossible.

The current problems surrounding construct validity violations in fMRI are especially insidious because they are effectively obscured by abstruse layers of statistics, math, and technical details associated with hemodynamic measurement. Furthermore, most researchers conducting neuroeconomics studies have backgrounds in economics and/or neuroscience, but tend to lack training in statistical methods that were developed over the last century for dealing with the measurement of complex, high-level, and sometimes ambiguous psychological processes. In essence, economists are now attempting to do psychology without dealing with the complicated issues of psychological inference that we have faced for over a century. We realize that, perhaps because psychology deals with everyday issues, research in psychology seems like it should be easy. But given the issues of validity we have discussed, along with the intensive methodological and statistical measures that must be taken to deal with the, the study of psychology is deceptively complicated.

We are concerned that economists, like many social neuroscientists, are currently in the process of reinventing the wheel when it comes to psychology. This is an issue because neuroeconomics has accumulated immense cachet in recent years, with a disproportionate number of recent publications in top journals such as *Science* and *Nature* and funding from national institutions. The concern is that most current findings from neuroeconomics will need to be revised as researchers gain a better understanding of brain structure and function, and much of the current literature will be obsolete. We suggest that more careful attention to psychology will improve the quality of the research and will preserve its utility for the science of human behavior.

Does this mean that neuroscience techniques are not of value to behavioral economics? Certainly not. Each is involved in a variety of research programs and is

yielding important new scientific knowledge. Similarly, the physiological measurement techniques underlying the lie detector contributed to our understanding of human functioning. Each technique holds out the potential to contribute to our understanding of human thought, feeling, and behavior. However, each also has the characteristics of the lie detector. They do not provide a clear "window into people's thoughts," and therefore we must interpret ambiguous information to understand their meaning. In our view, these limits simply bring physiology into line with verbal self-report and behavior observation as useful but flawed techniques.

A Unified Approach

We suggest, first, that the best approach to adopt is for researchers to focus on self-reported attitudes and feelings, observed behaviors, and physiological indicators at the same time and in the context of strong experimental designs. Together, they can provide converging validity for a construct of interest and for its experimental effects. We do not disparage any of these sources of information alone; we just think it is critical to consider all three. And, we think an important task is to map out their relationship to one another. Certainly, psychology should accept the centrality of behavior that is typical of studies by economists and should make a greater effort to link studies of subjective processes to behavioral outcomes.

Second, we think that issues of methodology need to be addressed directly, as is occurring today. Both psychologists and economists agree on the basic elements of an experimental method—for example, random assignment. And, we think that good science in both fields is based upon theory-driven hypothesis testing, rather than post hoc explanations for observed effects.

What differs? We would argue that psychologists deal more aggressively with the problems of the self-aware subject than do economists. We believe that this is an arena in which economists might benefit from greater consideration of the arguments of psychology. As it stands, economists appear to simply ignore the issue of self-presentation biases and, in some cases, assume that neuroimaging methods are immune to such biases. For example, in the studies done by anthropologists around the work involving the PDG and the dictator game, it does not seem to be an issue for economists to wonder what the participants were thinking as they played strange games with foreign social scientists. Rather it is assumed that their behavior reveals how they will act when they go to their local market to shop. Such assumptions about construct validity weaken the internal validity of an experiment and, as a direct result, drastically limit scientists' ability to draw meaningful interpretations from the data they collect. As economists begin to deal with these issues, they will find themselves faced with many of the problems that have led psychologists to think carefully about issues of validity and to address them through the judicious use of deception, indirect measures, and other methods designed to strengthen their inferences about the inner workings of the mind.

REFERENCES

Adolphs, R., D. Tranel, and A. R. Damasio. 1998. The Human Amygdala in Social Judgment. *Nature* **393**:470–474.

Akerloff, G. A. 2006. The Missing Motivation in Macroeconomics. Presidential address, American Economics Association, Chicago, IL.

Amodio, D. M. 2010. Can Neuroscience Advance Social Psychological Theory? Social Neuroscience for the Behavioral Social Psychologist. *Social Cognition* **28**:695–716.

Aronson, E., P. C. Ellsworth, J. M. Carlsmith, and M. H. Gonzalez. 1990. *Methods of Research in Social Psychology*. New York: McGraw-Hill.

Aronson, J. D. 2007. Brain Imaging: Culpability and the Juvenile Death Penalty. *Psychology, Public Policy, and Law* **13**:115–142.

Batson, D. 1991. *The Altruism Question: Toward a Social–Psychological Answer*. Hillsdale, NJ: Erlbaum.

Baumgartner, T., M. Heinrichs, A. Vonlanthen, U. Fischbacher, and E. Fehr. 2008. Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaption in Humans. *Neuron* **58**:639–650.

Cacioppo, J. T., G. G. Berntson, T. S. Lorig, C. J. Norris, E. Rickett, and H. Nusbaum. 2003. Just Because You're Imaging the Brain Doesn't Mean You Can Stop Using Your Head: A Primer and Set of First Principles. *Journal of Personality and Social Psychology* **85**:650–661.

Campbell, D. T. and J. Stanley. 1963. Experimental and Quasi-Experimental Designs for Research. In *Handbook of Research on Teaching*, ed. N. L. Gage. Chicago: Rand McNally.

Cronbach, L. J. and P. E. Meehl. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin* **52**:281–302.

Delgado, M. R., R. H. Frank, and E. A. Phelps. 2005. Perceptions of Moral Character Modulate the Neural Systems of Reward During the Trust Game. *Nature Neuroscience* **8**:1611–1618.

Engell, A. D., J. V. Haxby, and A. Todorov. 2007. Implicit Trustworthiness Decisions: Automatic Coding of Face Properties in Human Amygdala. *Journal of Cognitive Neuroscience* **19**:1508–1519.

Fehr, E. 2007. Human Nature and Social Cooperation. *Annual Review of Sociology* **33**:1–22.

Glimcher, P., E. Fehr, C. Camerer, and R. Poldrack. 2008. *Handbook of Neuroeconomics*. San Diego: Academic Press.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr. 2005. *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press.

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

Gollwitzer, P. M. 1990. Action Phases and Mind-sets. In *The Handbook of Motivation and Cognition: Foundations of Social Behavior*, eds. Higgins, E. T. and R. M. Sorrentino, vol. 2, 53–92. New York: Guilford Press.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreach. 2004. In *Foundations of Human Sociality*, eds. Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. New York: Oxford University Press.

Knutson, B., J. Taylor, M. Kaufman, R. Peterson, G. Glover. 2005. Distributed Neural Representation of Expected Value. *Journal of Neuroscience* **25**:4806–4812.

LaPiere, R. T. 1934. Attitudes vs. actions. *Social Forces* **13**:230–237.
McCabe, D. P. and A. D. Castel. 2008. Seeing is Believing: The Effect of Brain Images on Judgments of Scientific Reasoning. *Cognition* **107**:343–352.
Montague, P. R. and G. S. Berns. 2002. Neural Economics and the Biological Substrates of Valuation. *Neuron* **36**:265–284.
Plassmann H., J. O'Doherty, B. Shiv, and A. Rangel. 2008. Marketing Actions Can Modulate Neural Representations of Experienced Pleasantness. *Proceedings of the National Academy of Sciences* **105**:1050–1054.
Poldrack, R. A. 2006. Can Cognitive Processes Be Inferred from Neuroimaging Data? *Trends in Cognitive Sciences* **10**:59–63.
Sears, D.O. 1986. College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Psychology* **51**:515–530.
Tyler, T. R. 2006a. Social Motives and Institutional Design. In *The Evolution of Designed Institutions*, ed. G. V. Wangerheim. Blackwell.
Tyler, T. R. 2006b. Legitimacy and Legitimation. *Annual Review of Psychology* **57**:375–400.
Tyler, T. R. 2011. *Why People Cooperate*. Princeton, NJ: Princeton University Press.
Tyler, T. R. and S. Blader. 2000. *Cooperation in Groups: Procedural Justice, Social Identity, and Behavioral Engagement*. Philadelphia: Psychology Press.
Tyler, T. R. and D. DeCremer. 2006a. How Do We Promote Cooperation in Groups, Organizations, and Societies? The Interface of Psychology and Economics. In *Bridging Social Psychology*, ed. Paul van Lange. Philadelphia: Psychology Press.
Tyler, T. R. and D. DeCremer. 2006b. Cooperation in Groups. In *Social Psychology and Economics: Interdisciplinary Perspectives*, eds. D. DeCremer, M. Zeelenberg, and J. K. Murnighan. Mahwah, NJ: Erlbaum.
Webb, E. J., D. T. Campbell, D. Schwartz, and L. Sechrest. 1966. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand-McNally.
Weisberg, D., F. Keil, J. Goodstein, E. Rawson, and J. R. Gray. 2008. Illusions of Insight: The Curious Allure of Neuroscience Explanations. *Journal of Cognitive Neuroscience* **20**:470–477.
Yin, H. H. and B. J. Knowlton. 2006. The Role of the Basal Ganglia in Habit Formation. *Nature Reviews Neuroscience* **7**:464–476.

CHAPTER 12

THE HAMMER AND THE SCREWDRIVER

GARY CHARNESS

Is a hammer a better tool than a screwdriver? Sometimes one needs a hammer and sometimes one needs a screwdriver. They are different tools, suited for different purposes. Claiming superiority for one tool over the other seems misplaced. This principle also applies to research methods, as each method has its own strengths and weaknesses. This brings us to the current debate about the value of field experiments compared to the value of laboratory experiments. Levitt and List (2007) provide a well-known criticism of laboratory experiments, pointing out factors that are beneficial in field experiments while pointing out a number of factors that make the interpretation of data in laboratory experiments problematic. Their main issue is the degree to which "the insights gained in the lab can be extrapolated to the world beyond" (p. 153); this is also known as external validity. On the other hand, Falk and Heckman (2009) strongly emphasize the value of laboratory experiments. The thrust of their argument is that the controlled variation possible in laboratory experiments facilitates tests of theory, causal effects, and treatment effects.[1]

My own view is that laboratory experiments are best at testing theory and identifying treatment effects, and they can also provide useful qualitative insights. However, any assumption that the quantitative levels of behavior observed in the laboratory apply to naturally occurring settings must be carefully considered, as the laboratory is only a model of the field environment and cannot include many details that may influence behavior. Field experiments, for their part, offer promise in areas that are not readily susceptible to laboratory experimentation and generally involve a greater range of personal and demographic characteristics.

Field experiments are especially valuable to the extent that they can capture more realistic behavior (particularly in settings where the participants are unaware that there is an ongoing experiment). One should use the most appropriate tool or tools for the job at hand. Thus, I am a bit mystified at the heat in this debate over the relative value of field and laboratory experiments. In a certain sense, one wonders what all of the shouting is about.

Murningham (2008) discusses the idea that research methods are complementary rather than substitutes, pointing out that this is not a new idea. For example, a classic text in the social science is Runkel and McGrath (1972), who identify eight research strategies (including field studies, field experiments, and laboratory experiments), which they categorize along the two dimensions of obtrusive–unobtrusive research operations and universal–particular behavior.[2] They state: "We cannot emphasize too strongly our belief that none of these strategies has any natural or scientific claim to greater respect from researchers than any other" (p. 89). Runkel and McGrath (1972) advocate fitting the appropriate methodology to the research goals and circumstances. This would appear to be an eminently reasonable viewpoint.

On balance, field experiments can provide fairly good control of the environment, although rarely to the level attainable with laboratory experiments. Regarding the often-expressed notion that the participants in field studies are more representative of the relevant population, it is unclear whether undergraduate students are less representative than highly selected field populations such as sports-card traders, fruit pickers, bicycle messengers, tree planters, or school children. So it seems that neither field experiments nor laboratory experiments alone can capture a fully representative set of the population of interest.[3]

Despite the divergent views expressed in these articles and others, it is worth noting that there is indeed common ground. For example, in both Levitt and List (2007) and Falk and Heckman (2009), the authors discuss how one needs a model or theory to transport findings to new populations or environments, whether these data originate in laboratory experiments or field experiments. These articles also appear to agree that the controlled variation in the laboratory is better for careful tests of theory. Both state that there are shortcomings in both laboratory and field experiments but that each can provide useful insights. In fact, both camps apparently agree that both forms of experimentation (as well as hybrids) can be combined to yield a better understand of the phenomena involved.[4]

So, can't we all just get along? I might be trained to wield a hammer, but others may be experts with screwdrivers; in fact, I use both tools (field and laboratory experiments), as do other researchers. My sense is that all types of experimental practitioners are in the same boat. It behooves the crew to refrain from puncturing the hull.

NOTES

1. They state: "This control allows for the testing of precise predictions derived from game-theoretic models" (p. 636).

2. See the chart in Murningham (2008).

3. Furthermore, regarding the independence of observations in field experiments, Holt (2007) mentions a cautionary statistical point: "To the extent that social context and target demographics are important in a field experiment, each field experiment is in some sense like a data point that is specific to that combination of subjects and context unless appropriate random selection of subjects is employed" (p. 14).

4. One interesting point is that two of the key players in the debate (Falk and List) have used and continue to use both laboratory and field experiments in their research.

REFERENCES

Falk, A. and J. Heckman. 2009. Lab Experiments are a Major Source of Knowledge in the Social Sciences. *Science* **326**:535–538.

Holt, C. 2007. *Markets, Games, and Strategic Behavior*. Upper Saddle River, NJ: Prentice Hall.

Levitt, S. and J. List. 2007. What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *Journal of Economic Perspectives* **21**:153–174.

Murningham, J. K. 2008. A General Model for Experimental Inquiry in Economics and Social Psychology. Unpublished.

Runkel, P. and J. McGrath. 1972. *Research on Human Behavior: A Systematic Guide*. New York: Holt, Rinehart & Winston.

CHAPTER 13

DISCUSSION OF "PSYCHOLOGY AND ECONOMICS: AREAS OF CONVERGENCE AND DIFFERENCE"

THEO OFFERMAN

In their chapter, Tyler and Amodio note that although behavioral economists have moved toward studying "psychological constructs," they have refrained from updating their experimental method. Instead of obtaining direct measurements of phenomena as trust, altruism, and guilt aversion, economists continue to rely most heavily on observed behavior. Tyler and Amodio argue that it is time that economists start collecting direct evidence on these psychological constructs—for instance, by asking subjects to provide self-reports or by measuring some aspect of physiological or neurological activity. At the same time, they argue that psychologists should pay more attention to investigate whether the independently measured psychological constructs actually matter for human behavior.

Although Tyler and Amodio encourage economists to start collecting independent measures that clarify why people behave as they do, they warn against a naive embracement of using self-reports, physiological responses, and neuroscientific data. Self-reports may suffer from subjects' wish to present a desirable image of themselves. Neuroimaging and physiological measurements help to circumvent

the drawbacks of self-presentation biases, at least to a large extent. These methods introduce a new challenge when it comes to interpreting the data, however. Tyler and Amodio emphasize that complex psychological processes cannot directly be inferred from a pattern of brain activity or physiological responses.

I think that the chapter of Tyler and Amodio by and large provides a balanced view on the differences between economics and psychology. Their contribution is informative and thought provoking. Here, for the sake of discussion, I will focus on the issues where I disagree with the authors.

One central theme in the chapter concerns the awareness of subjects in experiments. If subjects are aware that they are studied, the danger is that they create a socially desirable impression of themselves. And alternatively, if subjects start guessing about what the experimenter's goal of the study is, they may behave in ways that they believe correspond to what the experimenter is looking for. They may even engage in the opposite behavior if they do not like the experimenter. In short, there is a danger of participants bias.

Tyler and Amodio propose two ways to diminish the danger of participants bias. One way is to proceed along the way of unobtrusive measurement. If people do not know that they are studied, they cannot alter their behavior in response to perceived experimenter demand effects. This solution is uncontroversial, but its practical usefulness may be limited because it is not always available. Tyler and Amodio state that the other way to prevent participants' bias lies in the use of deception. Probably most researchers who choose to employ deception do so because they think it helps to enhance the ecological validity of the experiment. According to Tyler and Amodio, a welcome byproduct of deception is that it diminishes participants bias. Their argument goes as follows. If the subjects are successfully led to believe that the experimenter is investigating question A instead of the actual question B, then participants will alter their behavior with respect to question A and act naturally with regard to question B. Thus, participants bias with regard to the actual question B is avoided.

Tyler and Amodio criticize economists for not dealing with participants bias. They illustrate their concerns with economists' investigations of social preferences. After initial work on unstructured bargaining, economists introduced structured games to get a sharper view on what motivates people when they make decisions in the social domain. To a large extent, stylized games like the trust game and ultimatum game dominated economists' thinking about social preferences in the 1980 and 1990. Noting that proposers might offer substantial amounts to responders in the ultimatum game because they correctly feared that trivial amounts would be rejected, the dictator game was designed. Then researchers realized that even the dictator game does not isolate subjects' motivation to give money. Allocators in a dictator game may be motivated to give part of their endowment by pure altruism, but they may also be driven by the fear for scorn of the receiver and/or experimenter. Therefore, in their relentless search to nail down subjects' true social

preferences, investigators designed a double-blind version of the dictator game where neither the experimenter nor the receiver could know who the allocator was.

Along the way, economists started collecting disturbing evidence on these simple games. Seemingly innocent features like framing or entitlements can have dramatic effects on the outcomes of simple stylized bargaining process (e.g., Hoffman and Spitzer, 1982, 1985; Guth and Tietz, 1986; Hoffman et al., 1994). For instance, subjects will bargain in a more selfish way when they feel that they have earned the right to act in the privileged role. So gradually it became clear that the quest to uncover people's stable social preferences was naive. Unfortunately, social preferences turn out to be an intricate phenomenon and context-dependent. Especially the outcomes of the most stylized bargaining game, the dictator game, were easily influenced by the experimenter. With retrospect, this may not be so surprising. It is hard to imagine that in an artificial game like the dictator game, subjects will not start wondering about what the experimenter is aiming for. In particular, when the experimenter uses a complicated procedure to make the experiment double blind, subjects may start guessing what is expected of them.

Ex post it is easy to criticize these experiments. However, if these experiments were not run, we would not know that social preferences are not the stable phenomenon that we had hoped for. Now it is time to stop running new variants of the dictator game and move on, however. My conclusion from this literature is that if we want to know how people bargain in a particular context, then it is important to get that particular context in the experimental design. Adding context has the advantage that the game becomes less artificial and that subjects will wonder less about the goal of the experimenter. In sufficiently rich bargaining experiments, subjects will even not be able to calculate the equilibrium predictions on the spot, and therefore they cannot behave according to the wishes of the experimenter even if they desire to do so (as, for instance, in Sonnemans et al. (2001) and Fréchette et al. (2003)).

It is also important to note that in most other economics experiments—for instance, on markets or auctions—subjects are not able to anticipate the theoretical predictions being tested. In the typical double auction experiment, a subject is only told about their own redemption values so that it is impossible to compute the competitive equilibrium. In the typical auction experiment, most human subjects are simply not able to derive the equilibrium predictions on the spot. It is reassuring that across nations, subjects' behavior in market games is more stable than in bargaining games (Roth et al., 1991).

I do not think that deception provides an adequate solution to participants bias. I agree that deception may help focus subjects' attention on a question other than the one being investigated. Deception introduces a new possible bias, however. The experimenter demand for the question not being investigated may be so strong that the subjects pay less attention to the actual question than they would have done in a natural setting. Take, for instance, the classic study by Darley and Batson (1973) on helping behavior. The experimenters instructed some subjects to give a speech

on the parable of the Good Samaritan and some others on a nonhelping relevant topic. To give the speech, subjects had to go to another building and on the way they passed a shabbily dressed person who was clearly in need of help. In the hurry condition, after subjects were instructed on the speech, the experimenter looked at his watch and said "Oh, your're late. They were expecting you a few minutes ago. We'd better get moving. The assistant should be waiting for you so you'd better hurry." In the low-hurry condition, subjects were told "It'll be a few minutes before they're ready for you, but you might as well head on over. If you have to wait over there, it shouldn't be long." Darley and Batson did not observe a significant difference in helping behavior between the subjects who were going to give a speech on the Good Samaritan and the ones with a speech on the nonhelping topic. They did observe a significant difference between those who were in a hurry and those who were not. The danger is that the subjects who were in a hurry condition were so eager to please the experimenter that they paid less attention to the person in need than they would have done in a natural hurry situation. As Milgram's experiments show, people have a very strong urge to obey the requests of an experimenter.

In many psychological experiments the experimenter and the subjects interact closely. Even when the experimenter disguises the goal of the experiment with deception, the danger of an opposite participants bias arises. By focusing subjects' attention on a decoy question, subjects may pay less attention to the real question than they would have done in a natural environment. In addition, the use of deception allows for a major drawback. There is a possibility that (experienced) subjects see through the deception and fool the experimenter instead of viceversa. Therefore, I think that deception should be avoided whenever possible. There are, of course, questions that can only be investigated with deception—take, for instance, Milgram's (1963) experimental procedure to investigate people's strong inclination to obey authority.

In my view, the solution chosen by economists is more promising. In the typical economics experiment, a substantial distance between the experimenter and the subjects is created. About 20 people sit behind computer screens in cubicles, and the experimenter does not walk around to look who is making what decision. Typically, subjects feel comfortable and unwatched in such a setup. Clear monetary incentives help override a possible remaining desire to behave in accordance with the wishes of the experimenter. And if one still fears that the incentives are not sufficient to accomplish this goal, there is the easy solution of enhancing the incentives. In psychological experiments, monetary incentives are usually absent. Participants' bias may be a bigger concern for the typical psychological experiment than for the typical economics experiment.

In their conclusion, Tyler and Amodio argue in favor of a unified approach, where researchers combine self-reported attitudes and feelings, physiological indicators, and observed behaviors to study the question of interest. So psychologists should maker greater efforts to link self-reports and physiological measures to actual behavior. And economists should start collecting direct measures of the

psychological constructs about which they are theorizing. Of course, it is hard to disagree with the proposition that such a unified approach will provide a fruitful testbed for theories on human behavior.

Still, I would not like to jump to the conclusion that all economists and psychologists should pursue the same unified approach. To the contrary, we may make the strongest scientific progress when the same problem is approached from different methodological angles. This way each discipline can make optimal use of its comparative advantage. Psychologists have a comparative advantage in thinking about clever measurements of psychological constructs. Economists have a comparative advantage in thinking through the consequences that changes in a treatment variable should have for behavior if subjects are truly motivated by phenomena as altruism or guilt aversion. The robustness of an empirical result can better be judged if researchers do not make use of exactly the same methodology. What is imperative is that researchers should not stop reading other people's work if it uses a different methodology. Social science may best be served by fruitful discussions between open-minded people who pursue different methods.

REFERENCES

Darley, J. M. and C. D. Batson. 1973. From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior. *Journal of Personality and Social Psychology* **27**:100–108.

Fréchette, G. R., J. H. Kagel, and S. F. Lehrer. 2003. Bargaining in Legislatures: An Experimental Investigation of Open Versus Closed Amendment Rules. *American Political Science Review* **97**:221–232.

Guth, W. and R. Tietz. 1986. Auctioning Ultimatum Bargaining Positions. In *Issues in West German Decision Research*, ed. R. W. Scholz. Frankfurt: Lang.

Hoffman, E., K. McCabe, K. Shachat, and V. Smith. 1994. Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior* **7**:346–380.

Hoffman, E. and M. Spitzer. 1982. The Coase Theorem: Some Experimental Tests. *Journal of Law and Economics* **25**:73–98.

Hoffman, E. and M. Spitzer. 1985. Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice. *Journal of Legal Studies* **15**:254–297.

Milgram, S. 1963. Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology* **67**:371–378.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. 1991. Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *American Economic Review* **81**:1068–1095.

Sonnemans, J., H. Oosterbeek, and R. Sloof. 2001. On the Relation Between Asset Ownership and Specific Investments. *Economic Journal* **111**:791–820.

PART IV

THE LABORATORY AND THE FIELD

WHAT DO LABORATORY EXPERIMENTS MEASURING SOCIAL PREFERENCES REVEAL ABOUT THE REAL WORLD?

STEVEN D. LEVITT
AND JOHN A. LIST

Economists have increasingly turned to the experimental model of the physical sciences as a method to understand human behavior. Peer-reviewed articles using the methodology of experimental economics were almost nonexistent until the mid-1960s and surpassed 50 annually for the first time in 1982; and by 1998, the number of experimental papers published per year exceeded 200 (Holt, 2006).

Steven D. Levitt is the Alvin H. Baum Professor of Economics and John A. List is Professor of Economics, University of Chicago, Chicago, Illinois. Levitt and List are both Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their e-mail addresses are ⟨s-levitt@uchicago.edu⟩ and ⟨jlist@uchicago.edu⟩, respectively.

Lab experiments allow the investigator to influence the set of prices, budget sets, information sets, and actions available to actors, and thus measure the impact of these factors on behavior within the context of the laboratory. The allure of the laboratory experimental method in economics is that, in principle, it provides *ceteris paribus* observations of individual economic agents, which are otherwise difficult to obtain.

A critical assumption underlying the interpretation of data from many laboratory experiments is that the insights gained in the lab can be extrapolated to the world beyond, a principle we denote as generalizability. For physical laws and processes like gravity, photosynthesis, and mitosis, the evidence supports the idea that what happens in the lab is equally valid in the broader world. The American astronomer Harlow Shapley (1964, p. 43), for instance, noted that "as far as we can tell, the same physical laws prevail everywhere." In this manner, astronomers are able to infer the quantity of certain gases in the Sunflower galaxy, for example, from observations of signature wavelengths of light emitted from that galaxy.

The basic strategy underlying laboratory experiments in the physical sciences and economics is similar, but the fact that humans are the object of study in the latter raises special questions about the ability to extrapolate experimental findings beyond the lab, questions that do not arise in the physical sciences. While few scientists would argue that observation influences whether Uranium_{239} would emit beta particles and turn into Neptunium, human behavior may be sensitive to a variety of factors that systematically vary between the lab and the outside world. In particular, we argue, based on decades of research in psychology and recent findings in experimental economics, that behavior in the lab is influenced not just by monetary calculations, but also by at least five other factors: 1) the presence of moral and ethical considerations; 2) the nature *and* extent of scrutiny of one's actions by others; 3) the context in which the decision is embedded; 4) self-selection of the individuals making the decisions; and 5) the stakes of the game. The remainder of this paper is devoted to examining how each of these factors influences decision making and the extent to which the environment constructed in the lab does or does not conform to real-world interactions on these various dimensions.[1]

We begin by presenting a simple model in which utility maximization is influenced not only by wealth maximization, but also by an individual's desire to "do the right thing" or make the "moral" choice. We then discuss the empirical evidence concerning the role of the five factors (above) in laboratory experiments.[2] Although our arguments apply more generally (Levitt and List, 2006), we focus the bulk of the discussion on the class of experiments that is believed to measure pro-social preferences. We provide a summary of the most popular games of this type in Table 1. We next discuss the extent to which the five factors systematically differ between laboratory experiments and naturally occurring environments, and explore how these differences affect the generalizability of experimental results outside the lab. We conclude that, just as is the case with naturally-occurring data, great caution

is required when attempting to generalize lab results out of sample: both to other populations and to other situations. Interpreting laboratory findings through the lens of theory helps us to understand the observed pattern of results and facilitates extrapolation of lab results to other environments. Field experiments, which exploit randomization in naturally-occurring settings, offer an attractive marriage of these competing empirical strategies.

A Model of Utility with Wealth and Morality

We begin by developing a model that makes precise our arguments regarding the potential factors that might influence individual decision-making. Many economists, dating back to Adam Smith, have emphasized that factors beyond wealth (for example, acting morally) enter into the utility function.[3] We do not claim originality in the ideas we are modeling. Rather, we view the model merely as a useful framework for organizing our discussion about the generalizability of results from laboratory experiments.

A utility-maximizing individual i is faced with a choice regarding a single action a. The choice of action affects the agent's utility through two channels. The first effect is on the individual's wealth (denoted W). The higher the stakes or monetary value of the game, which we denote v, the greater the decision's impact on W. The second effect is the nonpecuniary moral cost or benefit associated with the action, which we denote as M. Decisions which an individual views as immoral, antisocial, or at odds with his or her own identity (Akerlof and Kranton, 2000, 2005) may impose important costs on the decision maker (see also Gazzaninga, 2005). This moral payoff might vary across people, religions, or societies.

In practice, many factors influence the moral costs associated with an action, but for modeling purposes, we focus on just three aspects of the moral determinant. The first of these is the financial externality that an action imposes on others. The greater is the negative impact of an action on others, the more negative the moral payoff M. We model the externality as being an increasing function of the stakes of the game v. The second factor that influences the moral choice is the set of social norms or legal rules that govern behavior in a particular society. For instance, the mere fact that an action is illegal (for example, illicit drug use or smoking in restaurants), may impose an additional cost for partaking in such behavior. We denote these social norms against an action as n, with a greater value of n associated with a stronger norm against a behavior. Third, moral concerns depend on the nature and extent of how an individual's actions are scrutinized—such as whether the act is being televised, is taking place in front of one's children, or is performed

Table 1. Summary of Experimental Games Used to Measure Social Preferences

Name of game	Summary	Typical finding	Social preference interpretation
Ultimatum game[a]	A two-stage game where two people, a proposer and a responder, bargain over a fixed amount of money. In the first stage, the proposer offers a split of the money, and in the second stage, the responder decides to accept or reject the offer. If accepted, each player receives money according to the offer; if rejected, each player receives nothing.	*Proposer:* Majority of offers in the range of 25–50% of fixed amount. Few offers below 5%. *Responder:* Frequently reject offers below 20% of fixed amount.	*Proposer:* Fairness *Responder:* Punish unfair offers: negative reciprocity, fairness preferences, such as inequity aversion
Dictator game[b]	A variant of the ultimatum game: strategic concerns are absent as the proposer simply states what the split will be and the proposer has no veto power, rendering the proposed split as effective.	Usually more than 60% of subjects pass a positive amount of money, with the mean transfer roughly 20% of the endowment.	Altruism; fairness preferences, such as inequity aversion.
Trust game[c]	A sequential prisoner's dilemma game wherein the first mover decides how much money to pass to the second mover. All money passed is increased by a factor, $f > 1$, and the second mover then decides how much money to return to the first mover. In this light, the second mover is a dictator who has been given his endowment by the first mover.	*Proposer:* Average transfer of roughly 50% of endowment. *Responder:* Repayment is increasing in transfer. Average repayment rate is nearly 50% of transfer.	*Proposer:* Trust; foresee positive reciprocity *Responder:* Trustworthiness, positive reciprocity
Gift exchange game[d]	Similar to the trust game, but the money passed by the first mover (often labeled the "wage" or "price" offer), is not increased by a factor, rather it represents a pure lump-sum transfer. Also, the first mover requests a desired effort, or quality, level in return for the "wage" or "price" offer. The second mover then chooses an effort or quality level that is costly to provide, but increases the first mover's payoff.	*Proposer:* "Wage" or "price" offer is typically greater than the minimum allowed. *Responder:* Effort or quality increases in "wage" or "price" offer.	*Proposer:* Trust; foresee positive reciprocity *Responder:* Trustworthiness, positive reciprocity

Table 1. (continued)

Name of game	Summary	Typical finding	Social preference interpretation
Public goods game[e]	Generalization of the prisoner's dilemma game in that n group members decide simultaneously how much to invest in the public good. The payoff function is given by $P_i = e - g_i + \beta \Sigma_n g_j$, where e represents initial endowment; g_i is the level of tokens that subject i places in the group account; β is the marginal payoff of the public good; and $\Sigma_n g_j$ is the sum of the n individual contributions to the public good. By making $0 < \beta < 1 < n\beta$, the dilemma follows.	Players' contribution to public good is roughly 50% of endowment in one-shot games. Many players' contributions unravel to approach 0% in latter rounds of multi-period games	Altruism; fairness preferences, conditional reciprocity

[a] See Roth (1995) for a discussion of ultimatum and dictator games. This game was first proposed in the economics literature by Guth, Schmittberger, and Schwarze (1982).
[b] This game was first proposed in the economics literature by Kahneman, Knetsch, and Thaler (1986). A related game is the "punishment game," whereby an observer can, for a cost, punish the first mover by subtracting a portion of the first mover's payoff.
[c] This game was first proposed in the economics literature by Berg, Dickhaut, and McCabe (1995).
[d] This game was first proposed in the economics literature by Fehr, Kirchsteiger, and Riedl (1993), and a related game is Camerer and Weigelt (1988). The payoff function description for the buyer is similar to Fehr, Gachter, and Kirchsteiger's (1997) S13-S16 treatments. In this case, the price represents a pure lump-sum transfer, which differs from the earlier joint profit equation (Fehr, Kirchsteiger, and Riedl, 1993), which was characterized by price increases leading to an increase in the sum of payoffs when $q < 1$.
[e] See Ledyard (1995) for a discussion of these types games. This is a generalization of the famous prisoner's dilemma game.

under the watchful eye of an experimenter—as well as the way in which the process for reaching the decision and final allocation is emphasized (for example, in bargaining between husbands and wives, it is not just the final allocation that matters, but also the nature of the discussion by which the decision is reached). We denote the effect of scrutiny as s, with higher levels of s associated with greater moral costs.

With these considerations in place, focusing on the case in which utility is additively separable in the moral and wealth arguments, the utility function for individual i is:

$$U_i(a,\ v,\ n,\ s) = M_i(a,\ v,\ n,\ s) + W_i(a,\ v).$$

Framing the problem of utility maximization in this way yields several predictions. For example, in situations without a moral component, like the choice between investing in a stock or bond index, the model reverts back to a standard wealth maximization problem. However, when the wealth-maximizing action has a moral cost associated with it, the agent will deviate from that action to some extent towards an action that imposes a lower moral cost. The greater is the social norm against the wealth maximizing choice, or the greater the degree of scrutiny when the wealth-maximizing action has a social cost, the larger the deviation from that choice. In both cases, we envision the agent trading-off morality and wealth. When individuals follow different moral codes, they will generally make different choices when faced with the same decision problem. Typically, we expect that as the stakes of the game rise, wealth concerns will increase in importance relative to fairness concerns, although this need not always be the case.[4]

We would also expect that these various determinants of moral costs can interact, although the extent of such interaction remains an open empirical issue. For instance, for any given social norm n, as stakes v rise, the moral penalty for violating a given norm will be greater. As an example, people frown on shoplifting, but are much more forgiving of shoplifting than of embezzling millions of dollars. Likewise, the moral cost of violating a social norm increases as scrutiny s rises. For instance, an individual likely faces a larger utility loss from a crime if his capture is broadcast on CNN rather than merely recorded in his rap sheet.

The relevant social norms and the amount of scrutiny are not necessarily exogenously determined, but can be influenced in the context of real-world situations. For instance, panhandlers often emphasize physical deformities or carry placards claiming veteran's status to elicit greater sympathy from potential givers. When churches use "open" rather than "closed" collection baskets, they are acting in a manner consistent with recognition of the importance of norms and scrutiny, as potential contributors can not only see the total amount already gathered, but neighbors can witness each others' contributions (Soetevent, 2005).

The utility function we describe has relevance for a wide variety of behavior. For instance, the model explains why out-of-town visitors to a restaurant will leave a tip, even though they never intend to dine there in the future. Although leaving a tip imposes a financial cost on the diner, tipping provides an offsetting nonpecuniary reward. This behavior holds true even if one is eating alone, but probably even more so when there is a higher degree of scrutiny such as when you are accompanied by business clients, on a first date, or when another diner is observing your actions. Conlin, Lynn, and O'Donoghue (2003) present results from an extensive data set on tipping that confirms many of these intuitions.

Our primary interest here lies in developing the model's implications for the generalizability of lab experiments to naturally occurring contexts. When a laboratory experiment diverges from the real-world environment on certain dimensions of interest, the model provides a framework for predicting in what direction behavior in the lab will deviate from that outside the lab.

Implications for Experiments Designed To Measure Social Preferences

The issues that we raise are relevant for a wide range of experimental results, but their bite is likely to be greatest for those games with potential for a strong moral component to behavior. Research on social preferences, one of the most influential areas in experimental economics in recent years, fits this bill. This broad class of games, described earlier in Table 1, includes dictator and ultimatum bargaining games, public goods games, as well as trust and gift exchange games. Results from these types of experiments have been used to argue that pro-social preferences are important in a wide range of real-world settings (for example, Fehr and Gaechter, 2000; Camerer and Fehr, 2004)—an inference based on the assumption that the experimental findings are equally descriptive of the world at large.

In what follows, we examine the empirical evidence on possible complications arising when experimental findings are extrapolated to outside the lab. We are not denying that individuals have social preferences; indeed, our own model assumes that moral costs can be influenced by a concern for others as well as a concern for one's own appearance. Rather, we are interested in the extent to which the lab provides reasonable guidance as to the importance of such behavior in a wide range of naturally-occurring settings.

Scrutiny That Is Unparalleled in the Field

In the typical lab experiment, subjects enter an environment in which they are keenly aware that their behavior is being monitored, recorded, and subsequently scrutinized. Decades of research within psychology highlights the importance of the role obligations of being an experimental subject, the power of the experimenter herself, and the significance of the experimental situation. For instance, Orne (1962) wrote: "Just about any request which could conceivably be asked of the subject by a reputable investigator is legitimized by the quasi-magical phrase, 'This is an experiment,' and the shared assumption that a legitimate purpose will be served by the subject's behavior." Schultz (1969, p. 221) described the lab as having a "superior–subordinate" relationship matched only by that of "parent and child, physician and patient, or drill sergeant and trainee." Pierce (1908) warned of such effects almost a century ago:

> It is to the highest degree probable that the subject['s] . . . general attitude of mind is that of ready complacency and cheerful willingness to assist the investigator in every possible way by reporting to him those very things which he is most eager

> to find, and that the very questions of the experimenter. . . suggest the shade of reply expected. . . . Indeed . . . it seems too often as if the subject were now regarded as a stupid automaton.

The strength of such factors is so compelling that researchers in medical drug trials often go above and beyond using placebo and treatment groups by keeping the administrators themselves in the dark about which patients receive the treatment. In psychology experiments, to avoid demand-induced effects, subjects are often deceived about what exactly the investigator is measuring. In economics, however, deceptive practices are frowned upon. Clearly, the *nature* of scrutiny inherent in the lab is rarely encountered in the field, and represents an important aspect of the situation that needs to be accounted for when generalizing laboratory results.

Our theory suggests that such scrutiny will exaggerate the importance of prosocial behaviors relative to environments without such scrutiny. For example, List (2006) carries out gift exchange experiments in which buyers make price offers to sellers, and in return, sellers select the quality level of the good provided to the buyer. Higher quality goods are costlier for sellers to produce than lower quality goods, but are more highly valued by buyers. List began by running a standard gift exchange game in a laboratory context, but used experienced sports-card traders as subjects. The results mirrored the typical findings with other subject pools: strong evidence for social preferences was observed, in the sense that sellers offered higher quality levels to buyers who offered higher prices—although the sellers were not obligated by the rules of the game to do so. List then carried out a second lab experiment that maintained the central elements of the gift exchange game, but the goods exchanged in this lab treatment were actual baseball cards whose market values are heavily influenced by minor differences in condition that are difficult for untrained consumers to detect. If social preferences are present on the part of card sellers, then buyers who offer more money should be rewarded with higher-quality cards. When card sellers were brought into the lab to sell their cards, which were subsequently professionally graded, the results paralleled those obtained in the standard gift exchange game with student subjects.

List (2006) then moved from a lab environment (in which sellers knew their behavior was being scrutinized) to the sellers' natural environment. Importantly, dealers in this treatment were unaware that their behavior was being recorded as part of an experiment. Confederates were sent as buying agents to approach sellers on the floor of a sports-card show, instructing them to offer different prices in return for sports-cards of varying quality, just as in the lab treatment described above. When the dealers believed that consumers could not have the card graded or when there was likely to be little future interaction, little statistical relationship between price and quality emerged. Only when there were reputational consequences to a dealer (that is, when quality was verifiable and the potential for a long-term relationship existed), was high quality provided. The social preferences so routinely observed in the lab—even for this very same group of traders—were attenuated in the field.

Other field-generated data yield similar conclusions. For example, making use of personnel data from a leading United Kingdom–based fruit farm, Bandiera, Rasul, and Barankay (2005) find that behavior is consistent with a model of social preferences when workers can be monitored: when other workers can observe their productivity, workers internalize the negative externality that they impose on others under a relative compensation scheme. Yet this effect disappears when workers cannot monitor each other, which rules out pure altruism as the underlying cause of workers' behavior. Being monitored proves to be the critical factor influencing behavior in this study.

Relatedly, Benz and Meier (2006) compare how individuals behave in donation laboratory experiments with how the same individuals behave in the field. They find some evidence of correlation across situations, but find that subjects who have never contributed in the past to the charities gave 60 percent of their endowment to the charity in the lab experiment. Similarly, those who chose not to give to the charities in the two-year period after the experiment gave more than 50 percent of their experimental endowment to the charities in the lab experiment. Similar insights are reported in Laury and Taylor (forthcoming), who find little correlation between an "altruism parameter" estimated from a public goods lab experiment and actual contributions to a real public good (in this case, an urban tree-planting nonprofit organization).

In a "dining game," Gneezy, Haruvy, and Yafe (2004) find that behavior in a social dilemma game in the laboratory exhibits a considerable level of cooperative behavior—in the lab, students showed great reluctance to impose negative externalities. Yet, in a framed field experiment that resembles the laboratory game—diners were taken to eat at a restaurant—they find no evidence of cooperative play, even though both experimental samples are drawn from the same student population. They speculate that unfamiliarity with the task and confusion are two reasons why negative externalities are influential in the lab but not in the field. Such results are consistent with our model.

Overall, these results are consistent with the wealth of psychological literature that suggests there is only weak evidence of cross-situational consistency of behavior (for example, Mischel, 1968; Ross and Nisbett, 1991). Long ago, Hartshorne and May (1928) discovered that people who cheat in one situation are not the people who cheat in another. If this result spills over to measurement of pro-social preferences, it means either that (a) there is not a general cross-situational trait called "social preferences," and/or (b) the subjects view one situation as relevant to social preferences and the other as irrelevant.

Anonymity in the Lab and the Field

Another element of scrutiny is the degree of anonymity conferred upon experimental participants. Anonymity in this case takes two forms. One aspect of anonymity is between experimenter and subject; in some research designs, the experimenter

cannot determine what actions the subject takes. This aspect of anonymity is our primary focus. Additionally, there is the question of anonymity among subjects, an issue to which we devote less attention.

In the typical lab experiment, the identity of the subject can readily be linked to individual choices by the experimenter. Our theory predicts that the absence of anonymity will be associated with an increased level of pro-social behavior relative to settings in which individuals are more anonymous.

If the lack of anonymity between the experimenter and subject contributes to pro-social behaviors, then taking steps to reduce the extent to which subjects are knowingly observed should reduce the amount of such behavior.[5] To accomplish this goal, Hoffman et al. (1994; 1996) used a "double-blind" approach wherein the experimental monitor could not infer individual subjects' actions in a dictator game. Hoffman, McCabe, Shachat, and Smith (1994) find that 22 of 48 dictators (46 percent) donate at least $3 of a $10 pie under normal experimental conditions, but when subject–experimenter anonymity is added, only 12 of 77 dictators (16 percent) give at least $3. Hoffman, McCabe, Shachat, and Smith (1994, p. 371) conclude that observed "behavior may be due not to a taste for 'fairness' (other-regarding preferences), but rather to a social concern for what others may think, and for being held in high regard by others." Davis and Holt (1993, p. 269) note that these results "indicate that this apparent generosity is not altruism, but rather seems to arise largely from concerns about opinions of outside observers," which not only highlights the power of anonymity but also the important interaction between lab and anonymity effects. Consistent with this interpretation, Andreoni and Bernheim (2006) report subjects are much more likely to split the pie 50–50 in dictator games as scrutiny increases.[6]

List, Berrens, Bohara, and Kerkvliet (2004) adopt a different approach to generating anonymity between the subject and experimenter (as well as among subjects) using a "randomized response" technique. In this approach, for instance, a subject is told to answer "no" if either (a) they chose not to contribute to a public good, or (b) their mother was born in the first six months of the year. The experimenter therefore cannot determine with certainty whether the subject contributed to the public good or not. List, Berrens, Bohara, and Kerkvliet (2004) found that as decisions became less anonymous, a greater number of subjects opted to give to the public good in a one-shot decision. Both the degree of anonymity between the experimenter and subject, as well as anonymity among subjects, proved important.

Other dimensions of anonymity can also affect giving. For instance, Haley and Fessler (2005) find that giving in a dictator game significantly increases when a pair of eyes is shown on the computer screen where the dictator makes the allocation. This simple manipulation—meant to signal that the subjects' actions were being observed—increased the proportion of nonzero givers from 55 percent in the control treatment to 88 percent in the "eyespot" treatment. Likewise, Allen (1965) reports that increases in privacy reduce conformity. Individuals are also more likely to conform with the social norm of hand-washing when they are being observed (Harris and Munger, 1989).

Context Matters and Is Not Completely Controlled By the Experimenter

The actions people take are affected by a dazzlingly complex set of relational situations, social norms, frames, past experiences, and the lessons gleaned from those experiences. Consequently, the experimental investigator often lacks complete control over the full context within which the subject makes decisions (see also Harrison and List, 2004).

Experimentalists are fully aware that context in their instructions, inducing role-playing, framing, and the like can influence subject behavior (for example, Roth, 1995; Hertwig and Ortmann, 2001; Bohnet and Cooter, 2005). In a wide range of experimental settings, subtle manipulations have been shown to have drastic effects on actions. Rates of defection in prisoner dilemma games swing wildly depending on whether subjects are playing a "Community" or "Wall Street" game (Ross and Ward, 1996); using terms like "opponents" versus "partners" influences play in a myriad of games (Burnham, McCabe, and Smith, 2000, offer an example); asking people to "contribute" or to "allocate" funds in a linear public goods game matters, as does whether the game is framed as a positive externality or a negative one (Andreoni, 1995). Further, whether the agent "punishes" or "assigns" points to other agents can considerably influence play (for example, Gintis, 2001).

Contextual factors beyond the control of the experimenter appear to have equally profound impacts on actions. Henrich et al. (2005) provide powerful evidence of such effects. This group of scholars conducted one-shot ultimatum, dictator, and public goods games in 15 different small-scale communities in developing countries. They found enormous variation in behavior across communities, differences they were able to relate to patterns of everyday life and the social norms operating in these various communities. For instance, as Henrich et al. (2005, p. 31) note, the Orma readily recognized "that the public goods game was similar to the *harambee*, a locally-initiated contribution that Orma households make when a community decides to construct a public good such as a road or school," and subsequently gave quite generously. Likewise, among the whale-hunting Lamalera of Indonesia and the Ache in Paraguay, societies with strong norms of sharing, very generous ultimatum game offers are observed and very few offers are rejected. Alternatively, in small-scale foraging societies, such as the Hadza of Tanzania, low offers and high rejection rates are observed in ultimatum games. As Henrich et al. note (2005, p. 33) these "contrasting behaviors seem to reflect their differing patterns of life, not any underlying logic of hunter-gatherer life ways."

In all of the experiments Heinrich et al. (2005) conducted, the context that the experimenter can control—the payoffs, the description of the way the game is played, and so on—was almost identical. But the context that actors themselves

brought to the game and that experimenters cannot control—like past experiences and internalized social norms—proved centrally important in the outcome of play.

These examples highlight that an aspect of the lab over which experimenters have incomplete control is that subjects may not be playing the game that the experimenter intends. For instance, lab experiments in economics often seek to eliminate concerns as to whether behavior is motivated by a desire to build a reputation by using one-shot experimental designs. The basis for this methodology is that in a one-shot game, subjects will only display cooperative or pro-social behavior out of "social preference reciprocity," rather than because they are seeking to build and maintain a good reputation so other people will cooperate with them in the future. However, many real-world activities that have aspects of dictator, ultimatum, trust, or gift exchange games, public good provision, and other social dilemmas are typically not one-time encounters, but rather repeated games (for example, Hoffman, McCabe, and Smith, 1996; Ortmann and Hertwig, 2000; Harrison and Rutstrom, 2001). Effectively, personal experiences may cause the subjects to play these one-shot games as if they have some repetition, and the experimenter may have little or no ability to moderate this phenomenon. The Henrich et al. (2005) study of ultimatum games around the world showed that participants in laboratory games are likely to retrieve experiences and strategies that, unbeknownst to the experimenter, change the nature of the games. If an experimenter mistakenly assumes that the agent is treating the game as one-shot, reputation-building behavior can be misconstrued as social preferences.

While researchers might hope that experimental subjects will make clear strategic adjustments from repeated contexts to one-shot games, the empirical evidence is mixed. For instance, in a review of 15 studies that compare behavior across voluntary contribution games where subjects are randomly re-matched with new partners every round, as opposed to being paired with the same subjects over all rounds, Andreoni and Croson (forthcoming) report that five studies find more cooperation among the randomly re-matched, six find more cooperation among the repeatedly paired, and four studies fail to find a difference between the two treatments.

On the other hand, Fehr and Fischbacher (2003) find that responders react strongly to the possibility of acquiring a reputation; Andreoni and Miller (1993) report similar insights using data drawn from a prisoner's dilemma game. However, even results that suggest that subjects have an ability to distinguish between situations that have different prospects for future interaction do not necessarily imply that subjects behave in a one-shot experimental situation as if *no* prospect exists for future interaction. The received results are entirely consistent with a model whereby subjects recognize the difference between games with and without an explicit future, but still hold some prospect for future interaction in games described as one-shot (Samuelson, 2005).

While we know of no evidence that suggests those who exhibit strong social preferences in the lab behave similarly outside the lab, we do not doubt that such evidence can be collected. Yet, even if such data are gathered, many simple manipulations in the lab experiment can yield drastically different measures

of individual propensities. This result does not necessarily imply that preferences are labile. Rather, we view such data as evidence that when critical elements of the situation change, behavior will change in predictable ways.[7]

Stakes

Our model predicts that in games that have both a morality and wealth component, financial concerns will take on increasing prominence as the stakes rise. The evidence in the literature is only partially consistent with this view. In dictator games, a large increase in stakes generally leads to a less-than-proportionate increase in money transferred. For example, in Carpenter, Verhoogen, and Burks (2005), an increase in stakes from $10 to $100 caused the median offer to drop from 40 percent to 20 percent of the endowment. This result is much weaker for smaller changes in stakes: Cherry, Frykblom, and Shogren (2002) find no perceptible differences in offers across a $10 and $40 dictator game. Stakes effects have also been found in second-mover play in ultimatum games, in which the acceptance rate is generally increasing in the amount offered, conditional on the share offered—that is, a $1 offer in a $5 game is rejected more often than a $100 offer in a $500 game. Slonim and Roth (1998) find that in each range of offers below 50 percent, the acceptance rate goes up as the level of stakes increase (from 60 to 1500 Slovak koruna, the latter of which represents eight days of wages for the typical worker). In another type of game that involves some form of trust, the centipede game, Parco, Rapoport, and Stein (2002) similarly find that raising financial incentives causes a breakdown in mutual trust.[8] Fehr, Fischbacher, and Tougareva (2002), however, report fairness concerns play an important role for both low and high stakes in trust and gift exchange games.

We are not arguing that low stakes games in the lab have no market parallels; we take part in such transactions in well-functioning markets everyday. Our point is that if the analyst does not account properly for the differences in stakes across settings, inaccurate inference concerning the importance of pro-social preferences will likely result. The magnitude of such mismeasurement is a rich area for future research, and it would be interesting to compare the size of the low-stakes effect with that of the other factors discussed above.

Selection into the Experiment

If participants in laboratory studies differ in systematic ways from the actors engaged in the targeted real-world settings, attempts to generalize lab results directly might be frustrated. Most laboratory experiments have been conducted using students who self-select into the experiments. As Doty and Silverthorne (1975, p. 139) note, volunteers in human research "typically have more education, higher occupational status, earlier birth position, lower chronological age, higher need for

approval and lower authoritarianism than non-volunteers." Indeed, Rosenthal and Rosnow (1969) conclude that social experimentation is largely the science of "punctual college sophomore" volunteers, and have further argued that subjects are more likely to be "scientific do-gooders," interested in the research, or students who readily cooperate with the experimenter and seek social approval (see also Orne, 1962).[9]

In contrast, market participants are likely to be a selected sample of individuals whose traits allow them to excel in the marketplace. If such markets select agents who place a higher (or lower) value of W (or M) on decision tasks than student subjects, then one might suspect that the nature of the student selection into lab experiments might yield exaggerated pro-social behavior relative to such markets. On the other hand, lab participants may have less pro-social preferences than those who select into particular naturally-occurring environments, such as the clergy or public defenders.

One approach to investigating subject pool biases is to examine whether professionals, or other representative agents, and students behave similarly in laboratory experiments. Fehr and List (2004) examine experimentally how chief executive officers (CEOs) in Costa Rica behave in trust games and compare their behavior with that of Costa Rican students. They find that CEOs are considerably more trusting and exhibit more trustworthiness than students.[10] These differences in behavior may mean that CEOs are more trusting in everyday life, or it may be that CEOs are more sensitive to the lab and non-anonymity effects discussed above, or that the stakes are so low for the CEOs that the sacrifice to wealth of making the moral choice is infinitesimal.

A related issue concerns the possibility that only certain types of participants—students or professionals—are willing to take part in the experiment. For example, volunteers, whether students or CEOs, who have social preferences or who readily cooperate with the experimenter and seek social approval *might* be those who are most likely to participate in the experiment. In this case, games that purport to measure pro-social behaviors will yield upper bound estimates on the propensities of the target population.

Some limited but suggestive data from field and lab experiments supports this argument about selection into laboratory gift exchange experiments. When List (2006) approached a number of sports-card sellers about participating in the laboratory experiment described earlier, some sellers declined his invitation. But later and unbeknownst to them, these same sellers participated in the parallel field experiment. Those who declined to participate in the lab portion of the experiment were less pro-social in the field compared to dealers who agreed to participate in the lab experiment, although the differences were imprecisely measured due to small sample sizes and therefore not statistically significant at conventional levels. In a series of dictator games, Eckel and Grossman (2000) compare volunteers (those who select into the lab for an experiment) and pseudo-volunteers (those who are part of a class that is asked to participate during class time). Besides finding observable

differences across the subject pools, they find that pseudo-volunteers give more than volunteers, but also that volunteers behave in a less extreme manner than pseudo-volunteers.

Artificial Restrictions on Choice Sets and Time Horizons

Another issue closely related to those that we raise in the model is that in experiments, the researcher creates a set of rules governing the interactions, chooses the wording of instructions, and defines the set of actions the subject is allowed to take. In stark contrast, in naturally occurring environments, the choice set often is almost limitless and institutions arise endogenously.

Even among those who choose to participate in lab experiments, restrictions on the available choice set can affect observed behavior. For example, pro-social behavior might be observed less frequently in markets merely because people can avoid situations where they must make costly contributions to signal their generosity. This idea is illustrated in Lazear, Malmendier, and Weber (2006), who in an experiment allowed agents an opportunity to pay to opt out of playing the dictator game. They find that "the majority of subjects share without really wanting to, as evidenced by their willingness to avoid the dictator game and to even pay for avoiding it." Such forces are readily observable in the field as well—panhandlers receive less in gifts if passersby can easily "sort" themselves to the other side of the road to avoid interaction.

Another example of how the available choice set influences play in the dictator game can be found in Bardsley (2005) and List (forthcoming). In the typical dictator game, the subject is given, say, $10 and asked what portion the subject would like to share with the other player who received less than $10. The experiment is framed such that "giving nothing" is the least generous act, and substantial sums of money are given away. If instead, the subject is given $10 and is told that the rules allow giving any portion of this money away to the second player, or confiscating up to an additional $10 from the other player, subjects give little to the other player. Likewise, Andreoni, Brown, and Vesterlund (2002) make use of a sequential public goods game with an asymmetric equilibrium and find results consistent with the data in Bardsley and List. Real-world contexts typically offer the option of both giving and receiving, which may help explain in part why, contrary to the lab environment, people rarely receive anonymous envelopes with cash inside.

These examples also highlight that laboratory experiments often restrict the response mode to a single dimension, whereas real-world settings almost always involve multiple response modes. Consider again the act of giving within a dictator game. An agent who is inclined to help others might give money in the dictator game in the lab. In the field, this same agent might give nothing, instead using other

more efficient means to express generosity, such as volunteering time to help others. In this example, the laboratory evidence is consistent with some type of broader preference, but that preference might be expressed through a different activity in the field. Thus, when making comparisons across domains, one should take care to include all relevant dimensions.

Related to the choice set is the nature and temporal aspect of the task. Laboratory experiments usually consist of at most a few hours of fairly passive activities. For example, in a standard trust, or gift exchange, games in the laboratory, student subjects typically play several rounds of the game by choosing an effort or wage level (by circling or jotting down a number) in response to pecuniary incentive structures. The experiment usually lasts about an hour and a result often observed is that effort levels and wages are positively correlated. Such results are often interpreted as providing support for the received labor market predictions of Akerlof (1982) that the employer–employee relationship contains elements of gift exchange.

Such inference raises at least two relevant issues. First, is real-world, on-the-job effort different in nature from that required in lab tasks? Second, does the effect that we observe in the lab manifest itself over longer time periods? The evidence is sparse within the experimental economics literature on these issues, but studies are beginning to emerge. Using data gathered from a test of the gift exchange hypothesis in an actual labor market, Gneezy and List (2006) find that worker effort in the first few hours on the job is considerably higher in a "gift" treatment than in a "non-gift" treatment. However, after the initial few hours, no difference in outcomes was observed over the ensuing days of work. The notion that positive wage shocks do not invoke long-run effects in effort levels is also consistent with data reported in Al-Ubaydli, Steffen, Gneezy, and List (2006), Hennig-Schmidt, Rockenbach and Sadrieh's (2006) field experiment, and Kube, Maréchal, and Puppe (2006). These results suggest that great care should be taken before making inference from short-run laboratory experiments to long-run field environments.[11]

Such insights are consonant with results from the psychology literature in that important behavioral differences exist between short-run ("hot") and long-run ("cold") decision making. In the hot phase, visceral factors and emotions might prove quite important, whereas in the cold phase, immediate reactions may be suppressed. Loewenstein (2005) reviews some of the empirical evidence on behavioral differences across cold and hot states.

Generalizing the Findings of Laboratory Experiments to Actual Markets

We believe that several features of the laboratory setting need to be carefully considered before generalizing results from experiments that measure pro-social

behaviors to market settings they purport to describe. The model that we advance provides a framework to begin a discussion of the relevant economic and psychological factors that might influence behavior. Such factors include both the representativeness of the situation as well as the representativeness of the population: the nature and extent of scrutiny, the emphasis on the process by which decisions are made, the artificial limits placed on the action space, the imposition of task, the selection rules into the environments, and the stakes typically at risk.

In contrast to the lab, many real-world markets operate in ways that make pro-social behavior much less likely. In financial markets, for instance, the stakes are large, actors are highly anonymous, and little concern seems to exist about future analysis of one's behavior. Individuals with strong social preferences are likely to self-select away from these markets, instead hiring agents who lack such preferences to handle their financial dealings. Thus, one must take great care when claiming that patterns measured in the experimental economics laboratory are shared broadly by agents in certain real-world markets. It seems highly unlikely, for instance, that at the end of a day's trading, a successful trader would seek out the party that was on the wrong side of a market move and donate a substantial fraction of the day's profits to the person who lost—even though parallel behavior is routine in certain experiments. In addition, there is some trend in retail transactions away from an environment that fosters pro-social behavior towards one that does not, because of the rise of Internet sales and large retail chains.

In some naturally occurring settings, however, lab findings may *understate* the extent of pro-social actions. The degree of scrutiny present when making choices in front of one's children, or when one's actions are being televised, may far outstrip that in the lab. Thus, Levitt (2005) finds no evidence of discrimination towards blacks or women by participants on the televised game show "The Weakest Link." Also, inference from lab experiments measuring social preferences is typically based on interactions of complete strangers, anonymity between subjects, an absence of any social relations between subjects, and restricted communication channels between subjects. To the extent that such factors are not introduced into the lab environment by experimental subjects (yet, see Eckel and Wilson, 2004, footnote 15; Samuelson, 2005), such factors in the real world could induce a greater level of social preferences. For instance, one expects to find a great deal of altruism amongst family members, close friends, and comrades-in-arms. It is important to stress, however, that in settings with repeated interactions, it is difficult to distinguish between pro-social preferences and strategic actions taken with the goal of reputation building. Purely selfishly motivated individuals may forego short-term private gains, for instance, to support a cooperative equilibrium in an infinitely repeated prisoner's dilemma. When a firm treats an employee in a manner that is consistent with social preferences, the firm may simply be pursuing profit maximization. More careful empirical work in this area is warranted.

In addition, other important forces that are at work in naturally occurring markets can be absent in the lab. As Glaeser (2004) notes, it may be in the best interests of sophisticated agents to design institutions in such a way as to exploit the anomalous tendencies of others with whom they interact. Della Vigna and Malmendier (2006) provide an excellent example in the manner in which health clubs structure fees. Levitt (2004) similarly shows that bookmakers set lines that take advantage of the inherent biases of bettors. If certain markets are arranged whereby entrepreneurs must raise the prevalence of social behaviors to maximize their own profits, then the lab evidence might underestimate the importance of social preferences in comparison to such markets.[12]

Concluding Remarks

Perhaps the most fundamental question in experimental economics is whether findings from the lab are likely to provide reliable inferences outside of the laboratory. In this paper, we argue that the choices that individuals make depend not just on financial implications, but also on the nature and degree of others' scrutiny, the particular context in which a decision is embedded, and the manner in which participants are selected to participate. Because the lab systematically differs from most naturally occurring environments on these dimensions, experiments may not always yield results that are readily generalizable. Specifically, we argue that lab experiments generally exhibit a special type of scrutiny, a context that places extreme emphasis on the process by which decisions and allocations are reached, and a particular selection mechanism for participants. In contrast, many real-world markets are typified by a different type of scrutiny, little focus on process, and very different forms of self-selection of participants.

The points we make concerning generalizability of lab data apply with equal force to generalizing from data generated from naturally occurring environments. Empirical economists understand that studies of sumo wrestlers or sports-card traders cannot be seamlessly extrapolated to other economic settings. Any empirical estimate requires an appropriate theory for proper inference—and this lesson holds whether the data are obtained in the lab, from coin collector shows, or from government surveys. We envision similar practices among experimental economics: just as economists would want a model of firm and consumer behavior to tell what parameter we are estimating when we regress quantities on prices, we need a model of laboratory behavior to describe the data-generating process, and how it is related to other contexts. Theory is the tool that permits us to take results from one environment to predict in another, and generalizability of laboratory evidence should be no exception.

The discussion in this paper suggests three important conclusions regarding research design and interpretation. First, combining laboratory analysis with a model of decision-making, such as the model we present in this paper, expands

the potential role of lab experiments. By anticipating the types of biases common to the lab, experiments can be designed to minimize such biases. Further, knowing the sign and plausible magnitude of any biases induced by the lab, one can extract useful information from a study, even if the results cannot be seamlessly extrapolated outside the lab. In this sense, even in cases where lab results are believed to have little generalizability, some number from a laboratory estimate is better than no number, provided that a theoretical model is used to make appropriate inference.

Second, by adopting experimental designs that recognize the potential weaknesses of the lab, the usefulness of lab studies can be enhanced. For instance, one approach is to "nest" laboratory experiments one within another and then examine the different results of the related experiments. This approach may serve to "net out" laboratory effects and thus reveal more about deep structural parameters than running a simple, more traditional, experimental design. Additionally, lab experiments that focus on *qualitative* insights can provide a crucial first understanding and suggest underlying mechanisms that might be at work when certain data patterns are observed. Indeed, many of the arguments that we put forth in this study can be usefully explored using a laboratory experiment. Further, in the area of social dilemmas, laboratory experiments might help to illuminate whether punishing those who defect from pro-social behavior is a more powerful force than rewarding those who practice pro-social behavior.

Finally, recognizing that shortcomings exist in both lab-generated data and data from natural settings, an empirical approach that combines the best of each is appealing. A well-designed field experiment, incorporating the virtues of true randomization, but in a setting more representative of the behavior about which economists are seeking to learn, can serve as a bridge connecting these two empirical approaches.

Thanks to seminar participants at the 2005 International Meetings of the ESA for useful suggestions. Excellent suggestions from James Andreoni, Nicholas Bardsley, Gary Becker, Gary Charness, David Cooper, Dan Gilbert, Uri Gneezy, Hays Golden, Glenn Harrison, Reid Hastie, Dean Karlan, Dan Levin, Jayson Lusk, Ulrike Malmendier, Ted McConnell, Kevin Murphy, Andreas Ortmann, Charles Plott, Jesse Shapiro, Andrei Shleifer, Robert Slonim, and Richard Thaler greatly improved the study. Colin Camerer, Ernst Fehr, and Alvin Roth provided detailed comments and made suggestions that have resulted in many improvements, although not as many as they would have wished. Seminar participants at Brigham Young University, University of Chicago, Laval University, McMaster University, Oberlin College, and the University of Nevada, Reno also provided useful feedback on bits and pieces of this research. Financial support for Levitt came from the National Science Foundation and the Sherman Shapiro Research Fund. An earlier version of this paper that discussed the generalizability of a much wider class of experiments circulated under the title "What Do Laboratory Experiments Tell Us about the Real World?"

NOTES

1. There are instances where generalizability might not be of first-rate importance. For example, when testing a general theory, generalizability might not be a concern. In fact, as a first test of theory, an experimenter might wish to create an artificial environment for its own purpose: to create a clean test of the theory. Another example would be using the lab for methodological purposes—that is, to inform field designs by abstracting from naturally-occurring confounds.

2. This list certainly does not exhaust the set of reasons that lab experiments may not provide direct guidance with respect to behavior outside the lab. For instance, subjects tend to have less experience with the games they play in the lab, and there is no opportunity to seek advice from friends or experts in the lab. Also of potential importance is the fact that outside the lab, the design of institutions may be driven by sophisticated agents who seek ways to exploit the anomalous tendencies of those with whom they interact (Glaeser, 2004); this force is not at work inside the lab. For further discussion of these issues, see Harrison and List (2004) and Levitt and List (2006).

3. Smith viewed decisions as a struggle between "passions" and an "impartial spectator"—a "moral hector who, looking over the shoulder of the economic man, scrutinizes every move he makes" (Grampp, 1948, p. 317, as cited in Ashraf, Camerer, and Loewenstein, 2005). For formal models of such issues, see, for instance, Becker (1974), Akerlof (1982), and Bernheim (1994).

4. In Rabin's (1993) model, for sufficiently high stakes there will be no concern for fairness. Under our model there will be less, but potentially some, concern for others as stakes increase. Alternative models do exist, however; for example, Ledyard (1995) presents a model of voluntary contributions in which altruism and selfishness are traded off in such a way that an increase in the stakes has no influence on individual contributions.

5. We find the lab a fine tool to explore this type of scrutiny, yet manipulating the experimental environment in this manner may induce other difficulties in interpretation. For example, lessons learned from social psychologists teach us that such efforts to ensure anonymity might result in subjects inferring that the experimenter "demands" them to behave in a manner that might be deemed unacceptable (Loewenstein, 1999).

6. It should be noted, however, that Bolton, Zwick, and Katok (1998) and Laury and Taylor (1995) collect data that cast doubt on Hoffman, McCabe, Shachat, and Smith's (1994) results.

7. In this spirit, our arguments bear similarities to the Lucas critique.

8. The centipede game is an extensive form game that involves potentially several rounds of decisions. The game begins with player one deciding whether to take the payoff in the pot or to pass the decision to player two. If passed, player two then has a similar decision over a different payoff space. After each passing of the pot, the summation of payoffs is slightly increased, but the payoffs are arranged so that if one player passes and the opponent takes the pot, the player that passed receives less than if he or she had taken the pot.

9. When experimentally naïve high school students were asked, "How do you think the typical human subject is expected to behave in a psychology experiment?" over 70 percent circled characteristics labeled "cooperative" and "alert" (Rosenthal and Rosnow, 1973, pp. 136–7). However, these discussions typically revolve around social psychology experiments. Since economic experiments involve different subject matter and

involve monetary payments, such arguments might not generalize across disciplines. Kagel, Battalio, and Walker (1979) offer some evidence that volunteer subjects in an economics experiment have more interest in the subject than nonvolunteers, but other important variables are not different across volunteers and nonvolunteers.

10. Harbaugh, Krause, Liday, and Vesterlund (2003) conducted a set of trust experiments with students in third, sixth, ninth, and twelfth grade and found little variation across the participants in terms of trust and trustworthiness. However, in dictator games, the youngest children tend to make considerably smaller transfers than do older children and adults.

11. Naturally occurring data concerning the effects of pay shocks on work effort is mixed. Chen (2005), who uses a large data set drawn from the Australian Workplace Industrial Relations Survey to explore reciprocity in the workplace, finds little evidence consistent with reciprocity. Lee and Rupp (2006) examine the effort responses of U.S. commercial airline pilots following recent pay cuts, and find that such effects are very short-lived, consistent with Gneezy and List (2006). In the first week after a pay cut, frequent and longer flight delays are observed, but after the first week, airline flight performance reverts to previous levels. On the other hand, Krueger and Mas (2004) provide evidence consistent with negative reciprocity on the part of disgruntled Firestone employees, and Mas (forthcoming) documents persistent adverse effects on police performance following arbitration decisions in favor of the municipality.

12. In the long run, the impact of endogenously generated institutions on the amount of pro-social behavior is ambiguous. For instance, we learn from evolutionary biology that selection pressures can work against organisms that overextract from their hosts. In this case, firms that implement such policies can be displaced by firms that extract less from consumers. Even without such evolutionary competition, or in cases where incumbency advantages are large, if such institutions significantly raise the cost of faulty decision making, learning might occur more quickly and to a greater degree in markets than in the lab. However, if feedback mechanisms are weak in the field, such effects may generally not be observed.

REFERENCES

Akerlof, George A. 1982. "Labor Contracts as Partial Gift Exchange." *Quarterly Journal of Economics*, November, 97 (4):543–69.

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics*, August, 115(3): 715–53.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, Winter, 19(1): 9–32.

Allen, Vernon. 1965. "Situational Factors in Conformity." In *Advances in Experimental and Social Psychology*, vol. 2, ed. Leonard Berkowitz, 133–76. New York: Academic Press.

Al-Ubaydli, Omar, Steffen Andersen, Uri Gneezy, and John A. List. "Incentive Schemes to Promote Optimal Work Performance: Evidence from a Multi-Tasking Field Experiment." Unpublished paper, University of Chicago.

Andreoni, James. 1995. "Cooperation in Public Goods Experiments: Kindness or Confusion?" *American Economic Review*, 85(4): 891–904.

Andreoni, James, and B. Douglas Bernheim. 2006. "Social Image and the 50–50 Norm: Theory and Experimental Evidence." http://www.hss.caltech.edu/media/seminar-papers/bernheim.pdf.

Andreoni, James, Paul Brown, and Lise Vesterlund. 2002. "What Makes an Allocation Fair? Some Experimental Evidence." *Games and Economic Behavior*, 40(1): 1–24.

Andreoni, James, and Rachel Croson. Forthcoming. "Partners versus Strangers: Random Rematching in Public Goods Experiments." *Handbook of Experimental Economic Results*, ed. C. Plott and V. Smith.

Andreoni, James, and John Miller. 1993."Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence." *Economic Journal*, 103(418): 570–85.

Ashraf, Nava, Colin F. Camerer, and George Loewenstein. 2005. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives*, 19(3): 131–45.

Bandiera, Oriana, Iwan Rasul, and Imran Barankay. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics*, 120(3): 917–62.

Bardsley, Nicholas. 2005. "Altruism or Artifact? A Note on Dictator Game Giving." Center for Decision Research and Experimental Economics Discussion Paper 2005–10.

Becker, Gary S. 1974. "A Theory of Social Interactions." *Journal of Political Economy*, 82(6): 1063–93.

Benz, Matthias, and Stephan Meier. 2006."Do People Behave in Experiments as in Real Life? Evidence from Donations." Institute for Empirical Research in Economics, University of Zurich, Working Paper 248.

Berg, Joyce, John W. Dickhaut, and Kevin A. McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, July 10(1): 122–42.

Bernheim, B.D. 1994. "A Theory of Conformity." *Journal of Political Economy*, 102(5): 841–77.

Bohnet, Iris, and Robert D. Cooter. 2005. "Expressive Law: Framing or Equilibrium Selection?" John F. Kennedy School of Government Faculty Research Working Paper RWP03-046.

Bolton, Gary E., Rami Zwick, and Elena Katok. 1998. "Dictator Game Giving: Rules of Fairness Versus Acts of Kindness." *International Journal of Game Theory*, 27(2): 269–99.

Burnham, Terence, Kevin McCabe, and Vernon L. Smith. 2000. "Friend-or-Foe Intentionality Priming in an Extensive Form Trust Game." *Journal of Economic Behavior and Organization*, 43(1):57–73.

Camerer, Colin F., and Ernst Fehr. 2004 "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists." *In Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, eds. Joseph Henrich et al., 55–95. Oxford: Oxford University Press.

Camerer, Colin F., and Keith Weigelt. 1988. "Experimental Tests of a Sequential Equilibrium Reputation Model." *Econometrica*, 56(1): 1–36.

Carpenter, Jeffrey, Eric Verhoogen, and Stephen Burks. 2005. "The Effect of Stakes in Distribution Experiments." *Economics Letters*, March, 86(3): 393–8.

Chen, Paul. 2005. "Reciprocity at the Workplace: Do Fair Wages Lead to Higher Effort, Productivity, and Profitability?" http://gemini.econ.umd.edu/cgi-bin/conference/download.cgi?db_name=esam06&paper_id=222.

Cherry, Todd, Peter Frykblom, and Jason F. Shogren. 2002. "Hardnose the Dictator." *American Economic Review*, 92(4): 1218–21.

Conlin, Michael, Michael Lynn, and Ted O'Donoghue. 2003. "The Norm of Restaurant Tipping." *Journal of Economic Behavior and Organization*, 52(3): 297–321.

Davis, Douglas D., and Charles Holt. 1993. *Experimental Economics*. Princeton University Press.

DellaVigna, Stefano, and Ulrike Malmendier. 2006."Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719.

Doty, Richard L., and Colin Silverthorne. 1975. "Influence of Menstrual Cycle on Volunteering Behavior." *Nature*, (March, 13), 254(5496): 139–40.

Eckel, Catherine C., and Phillip Grossman. 1996. "Altruism in Anonymous Dictator Games." *Games and Economic Behavior*, 16(2): 181–191.

Eckel, Catherine C., and Philip J. Grossman. 2000. "Volunteers and Pseudo-Volunteers: The Effect of Recruitment Method in Dictator Experiments." *Experimental Economics*, 3(2): 107–120.

Eckel, Catherine C., and Rick K. Wilson. 2004. "Is Trust a Risky Decision?" *Journal of Economic Behavior and Organization*, December, 55(4): 447–65.

Fehr, Ernst, and Urs Fischbacher. 2003."The Nature of Human Altruism." *Nature*, 425(6960): 785–91.

Fehr, Ernst, Urs Fischbacher, and E. Tougareva. 2002. "Do High Stakes and Competition Undermine Fairness? Evidence from Russia." Institute for Empirical Research in Economics, University of Zürich, Working Paper 120.

Fehr, Ernst, and Simon Gaechter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159–81.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65(4): 833–60.

Fehr, Ernst, George Kirchsteiger, and Arno Riedl. 1993. "Does Fairness Prevent Market Clearing? An Experimental Investigation." *Quarterly Journal of Economics*, 108(2): 437–59.

Fehr, Ernst, and John A. List. 2004. "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOs." *Journal of the European Economic Association*, 2(5): 743–71.

Gazzaninga, Michael S. 2005. *The Ethical Brain*. Dana Press.

Gintis, Herbert. 2001. "The Contribution of Game Theory to Experimental Design in the Behavioral Sciences." *Behavioral and Brain Sciences*, 24(3): 411–12.

Glaeser, Edward. 2004. "Psychology and the Market." *American Economic Association Papers and Proceedings*, 94(2): 408–13.

Gneezy, Uri, Ernan Haruvy, and H. Yafe. 2004. "The Inefficiency of Splitting the Bill: A Lesson in Institution Design." *The Economic Journal*, April, 114(495): 265–80.

Gneezy, Uri, and John A. List. 2006. "Putting Behavioral Economics to Work: Field Evidence of Gift Exchange." *Econometrica*, September, 74(5): 1365–84.

Grampp, William. 1948. "Adam Smith and the Economic Man." *The Journal of Political Economy*, August, 56(4): 315–36.

Guth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization*, 3(2): 367–88.

Haley, Kevin J., and Daniel M. T. Fessler. 2005. "Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game." *Evolution of Human Behavior*, 26(3): 245–56.

Harbaugh, William T., Kate Krause, Steven G. Liday, and Lise Vesterlund. 2003. "Trust in Children." In *Trust, Reciprocity and Gains from Association: Interdisciplinary Lessons from Experimental Research*, ed. Elinor Ostrom and James Walker, 303–22. New York City, NY: Russell Sage Foundation.

Harris, S. J., and K. Munger. 1989. "Effects of an Observer on Hand Washing in Public Restrooms." *Perceptual and Motor Skills*, 69, pp. 733–5.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature*, December, 42(4): 1009–55.

Harrison, Glenn W., and Elisabet Rutstrom. 2001. "Doing It Both Ways—Experimental Practice and Heuristic Context." *Behavioral and Brain Sciences*, 24(3): 413–4.

Hartshone, Hugh, and Mark A. May. 1928. *Studies in Deceit*. New York: Macmillan.

Hennig-Schmidt, Heike, Bettina Rockenbach, and Abdolkarim Sadrieh. 2006. "In Search of Workers' Real Effort Reciprocity—A Field and a Laboratory Experiment." Governance and the Efficiency of Economic Systems Working Paper 55. Free University of Berlin, Humboldt University of Berlin.

Henrich, Joseph., et al. 2005. "Economic Man" in Cross-Cultural Perspective: Ethnography and Experiments from 15 Small-Scale Societies." *Behavioral and Brain Sciences*. 28(6): 795–815.

Hertwig, Ralph, and Andreas Ortmann. 2001."Experimental Practices in Economics: A Challenge for Psychologists?" *Behavioral and Brain Sciences*, 24(4): 383–451.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon 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*, June, 86(3): 653–60.

Holt, Charles A. 2006. *Markets, Games, and Strategic Behavior: Recipes for Interactive Learning*. Addison-Wesley.

Kagel, John H., Raymond C. Battalio, and James M. Walker. 1979. "Volunteer Artifacts in Experiments in Economics: Specification of the Problem and Some Initial Data from a Small-Scale Field Experiment." *Research in Experimental Economics*, ed. Vernon L. Smith, 169–97. JAI Press.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review*, 76(4): 728–41.

Krueger, Alan B., and Alexandre Mas. 2004. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy*, 112(2): 253–89.

Kube, Sebastian, Michel André Maréchal, and Clemens Puppe. 2006. "Putting Reciprocity to Work: Positive versus Negative Responses in the Field." University of St. Gallen Department of Economics Discussion Paper 2006–27.

Laury, Susan K., James M. Walker, and Arlington W. Williams. 1995. "Anonymity and the Voluntary Provision of Public Goods." *Journal of Economic Behavior and Organization*, 27(3): 365–80.

Laury, Susan K., and Laura O. Taylor. Forthcoming. "Altruism Spillovers: Are Behaviors in Context-free Experiments Predictive of Altruism toward a Naturally Occurring Public Good?" *Journal of Economic Behavior and Organization*.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2006. "Sorting in Experiments." National Bureau of Economic Research Working Paper 12041.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, ed. J. Kagel and A. Roth, chap. 2. NJ: Princeton University Press.

Lee, Darin, and Nicholas G. Rupp. 2006. "Retracting a Gift: How Does Employee Effort Respond to Wage Reductions?" http://www.ecu.edu/econ/wp/05/ecu0523.pdf.

Levitt, Steven D. 2004. "Why Are Gambling Markets Organized So Differently from Financial Markets?" *Economic Journal*, 114(495): 223–46

Levitt, Steven D. 2005. "Testing Theories of Discrimination: Evidence from the Weakest Link." *Journal of Law & Economics*, 47(2): 431–52.

Levitt, Steven D., and John A. List. 2006. "What Do Laboratory Experiments Tell Us about the Real World?" http://pricetheory.uchicago.edu/levitt/Papers/jep%20revision%20Levitt%20&%20List.pdf.

List, John A. 2006. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *Journal of Political Economy*, 114(1): 1–37.

List, John A. Forthcoming. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy*.

List, John A., Robert Berrens, Alok Bohara, and Joe Kerkvliet. 2004. "Examining the Role of Social Isolation on Stated Preferences." *American Economic Review*, 94(3), pp. 741–52.

Loewenstein, George. 1999. "Experimental Economics from the Vantage-Point of Behavioural Economics." *Economic Journal*, February 109(453): F23–34

Loewenstein, George. 2005. "Hot-cold Empathy Gaps and Medical Decision-Making." *Health Psychology*, 24(4): S49–S56.

Mas, Alexandre. 2006. "Pay, Reference Points, and Police Performance." *Quarterly Journal of Economics*, 121(3): 783–821.

Mischel, Walter. 1968. *Personality and Assessment*. New York; Wiley.

Orne, Martin T. 1962. "On the Social Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist*, 17(10): 776–83.

Ortmann, Andreas, and Ralph Hertwig. 2000."Why Anomalies Cluster in Experimental Tests of One-Shot and/or Finitely Repeated Games: Some Evidence from Psychology." http://home.cerge.cuni.cz/Ortmann/Papers/10baseratesii06142000.pdf.

Parco, James E., Amnon Rapoport, and William E. Stein. 2002. "Effects of Financial Incentives on the Breakdown of Mutual Trust." *Psychological Science*, 13(3): 292–7.

Pierce, A. H. 1908. "The Subconscious Again." *Journal of Philosophy, Psychology, & Scientific Methods*, 5(10): 264–71.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, December, 83(5): 1281–1302.

Rosenthal, Robert W., and Ralph L. Rosnow. 1969. *Artifact in Behavioral Research*. New York: Academic Press.

Rosenthal, Robert W., and Ralph L. Rosnow. 1973. *The Volunteer Subject*. New York: John Wiley and Sons.

Ross, Lee, and Andrew Ward. 1996."Naive Realism: Implications for Social Conflict and Misunderstanding." In *Values and Knowledge*, ed. T. Brown, E. Reed, and E. Turiel, 103–35. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ross, Lee, and Richard E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. New York: McGraw-Hill.

Roth, Alvin E. 1995. "Bargaining Experiments." In *The Handbook of Experimental Economics*, ed. J. H. Kagel and E. R. Alvin, 253–342. Princeton, NJ: Princeton University Press.

Samuelson, Larry. 2005. "Economic Theory and Experimental Economics." *Journal of Economic Literature*, March, 43(1): 65–107.

Schultz, Duane P. 1969. "The Human Subject in Psychological Research." *Psychological Bulletin*, 72(3): 214–28.

Shapley, Harlow. 1964. *Of Stars and Men: Human Response to an Expanding Universe*. Westport CT: Greenwood Press.

Slonim, Robert and Alvin E. Roth. 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." *Econometrica*, 66(3), pp. 569–96.

Soetevent, Adriaan R. 2005. "Anonymity in Giving in a Natural Context—A Field Experiment in Thirty Churches." *Journal of Public Economics*, 89(11–12): 2301–23, Table 1.

This article has been cited by

1. Jetske A. Bouma, K.J. Joy, Suhas Paranjape, Erik Ansink. 2014. The Influence of Legitimacy Perceptions on Cooperation – A Framed Field Experiment. *World Development* **57**, 127–137.
2. Nicola Bellantuono, Donatella Ettorre, Gregory E. Kersten, Pierpaolo Pontrandolfo. 2014. Multi- attribute Auction and Negotiation for e-Procurement of Logistics. *Group Decision and Negotiation* **23**:3, 421–441.
3. Dan Silverman, Joel Slemrod, Neslihan Uler. 2014. Distinguishing the role of authority "in" and authority "to". *Journal of Public Economics* **113**, 32–42.
4. Guy Grossman and Delia Baldassarri. 2012. The impact of elections on cooperation: Evidence from a Lab-in-the-Field Experiment in Uganda. *American Journal of Political Science* **56**:4, 964–985.
5. Marco Paccagnella, Paolo Sestito. 2014. School cheating and social capital. *Education Economics* 1–22.
6. Maren Elise Bachke, Frode Alfnes, Mette Wik. 2014. Eliciting Donor Preferences. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations* **25**:2, 465–486.
7. Brenna Ellison, Jayson L. Lusk, David Davis. 2014. The Impact Of Restaurant Calorie Labels on Food Choice: Results from a Field Experiment. *Economic Inquiry* **52**:2, 666–681.
8. Robin M. Hogarth, Marie Claire Villeval. 2014. Ambiguous incentives and the persistence of effort: Experimental evidence. *Journal of Economic Behavior & Organization* **100**, 1–19.

9. Guang-Xin Xie, Robert Madrigal, David M. Boush. 2014. Disentangling the Effects of Perceived Deception and Anticipated Harm on Consumer Responses to Deceptive Advertising. *Journal of Business Ethics*.
10. Saraï Sapulete, Arjen van Witteloostuijn, Wesley Kaufmann. 2014. An experimental study into the influence of works council advice on managerial decision-making. *Scandinavian Journal of Management*.
11. Ayana Elizabeth Johnson, Daniel Kaiser Saunders. 2014. Time preferences and the management of coral reef fisheries. *Ecological Economics* **100**, 130–139.
12. Johannes Abeler, Daniele Nosenzo. 2014. Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*.
13. Alex Shaw, Kristina Olson. 2014. Fairness as partiality aversion: The development of procedural justice. *Journal of Experimental Child Psychology* **119**, 40–53.
14. Juan Camilo Cárdenas, Nicolas Roux, Christian R. Jaramillo, 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.
15. Karen Croxson, J. James Reade. 2014. Information and Efficiency: Goal Arrival in Soccer Betting. *The Economic Journal* **124**:575, 62–91.
16. Petr Sauer. 2014. Art of Designing Teaching Laboratory Experiments: The Case of Water Management. *Procedia - Social and Behavioral Sciences* **122**, 204–209.
17. Brent J Davis, David B Johnson. 2014. Water Cooler Ostracism: Social Exclusion as a Punishment Mechanism. *Eastern Economic Journal*.
18. Philipp Schreck. 2014. Honesty in managerial reporting: How competition affects the benefits and costs of lying. *Critical Perspectives on Accounting*.
19. Magali A. Delmas, Neil Lessem. 2014. Saving power to conserve your reputation? The effectiveness of private versus public information. *Journal of Environmental Economics and Management*.
20. Darwin Choi, Sam K. Hui. 2014. The role of surprise: Understanding overreaction and underreaction to unanticipated events using in-play soccer betting market. *Journal of Economic Behavior & Organization*.
21. Till Proeger, Lukas Meub. 2014. Overconfidence as a social bias: Experimental evidence. *Economics Letters* **122**:2, 203–207.
22. Philipp Doerrenberg, Denvil Duncan. 2014. Experimental evidence on the relationship between tax evasion opportunities and labor supply. *European Economic Review*.
23. Jonathan E. Alevy, Francis L. Jeffries, Yonggang Lu. 2014. Gender- and frame-specific audience effects in dictator games. *Economics Letters* **122**:1, 50–54.
24. Gary E. Bolton, Axel Ockenfels. 2014. Does laboratory trading mirror behavior in real world markets? Fair bargaining and competitive bidding on eBay. *Journal of Economic Behavior & Organization* **97**, 143–154.
25. Maria Börjesson, Jonas Eliasson. 2014. Experiences from the Swedish Value of Time study. *Transportation Research Part A: Policy and Practice* **59**, 144–158.
26. Agnès Festré, Pierre Garrouste. 2014. Theory and Evidence in Psychology and Economics about Motivation Crowding Out: A Possible Convergence?. *Journal of Economic Surveys* n/a-n/a.
27. Kristen Hawkes. 2013. Primate Sociality to Human Cooperation. *Human Nature*.
28. William T. Harbaugh, Naci Mocan, Michael S. Visser. 2013. Theft and Deterrence. *Journal of Labor Research* **34**:4, 389–407.

29. Hannes Koppel, Günther G. Schulze. 2013. The Importance of the Indirect Transfer Mechanism for Consumer Willingness to Pay for Fair Trade Products—Evidence from a Natural Field Experiment. *Journal of Consumer Policy* **36**:4, 369–387.
30. Olivier Armantier, Amadou Boly. 2013. Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada. *The Economic Journal* **123**:573, 1168–1187.
31. Paola Sapienza, Anna Toldra-Simats, Luigi Zingales. 2013. Understanding Trust. *The Economic Journal* **123**:573, 1313–1332.
32. Victor Iajya, Nicola Lacetera, Mario Macis, Robert Slonim. 2013. The effects of information, social and financial incentives on voluntary undirected blood donations: Evidence from a field experiment in Argentina. *Social Science & Medicine* **98**, 214–223.
33. Danae Manika, Victoria K. Wells, Diana Gregory-Smith, Michael Gentry. 2013. The Impact of Individual Attitudinal and Organisational Variables on Workplace Environmentally Friendly Behaviours. *Journal of Business Ethics*.
34. Sigbjørn Birkeland, Alexander W. Cappelen, Erik Ø. Sørensen, Bertil Tungodden. 2013. An experimental study of prosocial motivation among criminals. *Experimental Economics*.
35. Gary Charness, David Masclet, Marie Claire Villeval. 2013. The Dark Side of Competition for Status. *Management Science* 131105054351001.
36. Helen Z. Margetts, Peter John, Scott A. Hale, Stéphane Reissfelder. 2013. Leadership without Leaders? Starters and Followers in Online Collective Action. *Political Studies* n/a-n/a.
37. E.N. Speelman, L.E. García-Barrios, J.C.J. Groot, P. Tittonell. 2013. Gaming for smallholder participation in the design of more sustainable agricultural landscapes. *Agricultural Systems*.
38. Vittorio Pelligra, Luca Stanca. 2013. To give or not to give? Equity, efficiency and altruistic behavior in an artefactual field experiment. *The Journal of Socio-Economics* **46**, 1–9.
39. Susanne Neckermann, Bruno S. Frey. 2013. And the winner is…? The motivating power of employee awards. *The Journal of Socio-Economics* **46**, 66–77.
40. Hanna J. Ihli, Syster C. Maart-Noelck, Oliver Musshoff. 2013. Does timing matter? A real options experiment to farmers' investment and disinvestment behaviours. *Australian Journal of Agricultural and Resource Economics* n/a-n/a.
41. C. Ehm, C. Kaufmann, M. Weber. 2013. Volatility Inadaptability: Investors Care About Risk, but Cannot Cope with Volatility. *Review of Finance*.
42. Rudolf Vetschera, Guenther Kainz. 2013. Do Self-Reported Strategies Match Actual Behavior in a Social Preference Experiment?. *Group Decision and Negotiation* **22**:5, 823–849.
43. Philip J. Cash, Ben J. Hicks, Steve J. Culley. 2013. A comparison of designer activity using core design situations in the laboratory and practice. *Design Studies* **34**:5, 575–611.
44. Anna Dreber, Tore Ellingsen, Magnus Johannesson, David G. Rand. 2013. Do people care about social context? Framing effects in dictator games. *Experimental Economics* **16**:3, 349–371.
45. Blair L. Cleave, Nikos Nikiforakis, Robert Slonim. 2013. Is there selection bias in laboratory experiments? The case of social and risk preferences. *Experimental Economics* **16**:3, 372–382.
46. Amine Ouazad, Lionel Page. 2013. Students' perceptions of teacher biases: Experimental economics in schools. *Journal of Public Economics* **105**, 116–130.

47. Ryan Bubb, Alex Kaufman. 2013. Consumer biases and mutual ownership. *Journal of Public Economics* **105**, 39–57.
48. Marcela Ibanez, Peter Martinsson. 2013. Curbing coca cultivation in Colombia — A framed field experiment. *Journal of Public Economics* **105**, 1–10.
49. Carl D. Mildenberger. 2013. The constitutional political economy of virtual worlds. *Constitutional Political Economy* **24**:3, 239–264.
50. Christina Gravert. 2013. How luck and performance affect stealing. *Journal of Economic Behavior & Organization* **93**, 301–304.
51. Sebastian Lotz, Thomas Schlösser, Daylian M. Cain, Detlef Fetchenhauer. 2013. The (in)stability of social preferences: Using justice sensitivity to predict when altruism collapses. *Journal of Economic Behavior & Organization* **93**, 141–148.
52. Michael Pickhardt, Aloys Prinz. 2013. Behavioral dynamics of tax evasion – A survey. *Journal of Economic Psychology*.
53. Roger Hartley, Gauthier Lanot, Ian Walker. 2013. Who Really Wants to be a Millionaire? Estimates Of Risk Aversion From Gameshow Data. *Journal of Applied Econometrics* n/a-n/a.
54. J. F. Suter, C. A. Vossler. 2013. Towards an Understanding of The Performance of Ambient Tax Mechanisms in The Field:Evidence from Upstate New York Dairy Farmers. *American Journal of Agricultural Economics*.
55. Cosmina Bradu, Jacob L. Orquin, John Thøgersen. 2013. The Mediated Influence of a Traceability Label on Consumer's Willingness to Buy the Labelled Product. *Journal of Business Ethics*.
56. Hans-Theo Normann, Till Requate, Israel Waichman. 2013. Do short-term laboratory experiments provide valid descriptions of long-term economic interactions? A study of Cournot markets. *Experimental Economics*.
57. Sebastian Kube, Michel André Maréchal, Clemens Puppe. 2013. Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment. *Journal of the European Economic Association* **11**:4, 853–870.
58. Armin Falk, Stephan Meier, Christian Zehnder. 2013. Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples. *Journal of the European Economic Association* **11**:4, 839–852.
59. Menusch Khadjavi, Andreas Lange. 2013. Prisoners and their dilemma. *Journal of Economic Behavior & Organization* **92**, 163–175.
60. Gregory Gurevich, Doron Kliger. 2013. The Manipulation: Socio-Economic Decision Making. *Journal of Economic Psychology*.
61. Toby Bolsen, Paul J. Ferraro, Juan Jose Miranda. 2013. Are Voters More Likely to Contribute to Other Public Goods? Evidence from a Large-Scale Randomized Policy Experiment. *American Journal of Political Science* n/a-n/a.
62. Michelle Jackson, D.R. Cox. 2013. The Principles of Experimental Design and Their Application in Sociology. *Annual Review of Sociology* **39**:1, 27–49.
63. Daniel Alfredo Revollo-Fernandez, Alonso Aguilar-Ibarra. 2013. Measures of risk associated to regulations compliance: a laboratory experiment on the use of common-pool resources. *Journal of Risk Research* 1–19.
64. Thierry Madiès, Marie Claire Villeval, Malgorzata Wasmer. 2013. Intergenerational attitudes towards strategic uncertainty and competition: A field experiment in a Swiss bank. *European Economic Review* **61**, 153–168.
65. Jeffrey Winking, Nicholas Mizer. 2013. Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior* **34**:4, 288–293.

66. Hakan J. Holm, Sonja Opper, Victor Nee. 2013. Entrepreneurs Under Uncertainty: An Economic Experiment in China. *Management Science* **59**:7, 1671–1687.
67. E. Yoeli, M. Hoffman, D. G. Rand, M. A. Nowak. 2013. Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences* **110**:Supplement_2, 10424–10429.
68. Jan Stoop. 2013. From the lab to the field: envelopes, dictators and manners. *Experimental Economics*.
69. Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* **47**:4, 2025–2047.
70. Jon Anderson, Stephen V. Burks, Jeffrey Carpenter, Lorenz Götte, Karsten Maurer, Daniele Nosenzo, Ruth Potter, Kim Rocha, 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, 170–189.
71. Fredrik Carlsson, Haoran He, Peter Martinsson. 2013. Easy come, easy go. *Experimental Economics* **16**:2, 190–207.
72. Pamela Jakiela. 2013. Equity vs. efficiency vs. self-interest: on the use of dictator games to measure distributional preferences. *Experimental Economics* **16**:2, 208–221.
73. Axel Franzen, Sonja Pointner. 2013. The external validity of giving in the dictator game. *Experimental Economics* **16**:2, 155–169.
74. Robert Slonim, Carmen Wang, Ellen Garbarino, Danielle Merrett. 2013. Opting-in: Participation bias in economic experiments. *Journal of Economic Behavior & Organization* **90**, 43–70.
75. Patrick McAlvanah, Charles C. Moul. 2013. The house doesn't always win: Evidence of anchoring among Australian bookies. *Journal of Economic Behavior & Organization* **90**, 87–99.
76. Erin L. Krupka, Roberto A. Weber. 2013. Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?. *Journal of the European Economic Association* **11**:3, 495–524.
77. Terence C. Burnham. 2013. Toward a neo-Darwinian synthesis of neoclassical and behavioral economics. *Journal of Economic Behavior & Organization* **90**, S113-S127.
78. Fabrice Etilé, Sabrina Teyssier. 2013. Corporate social responsibility and the economics of consumer social responsibility. *Revue d'Études en Agriculture et Environnement* **2013**:02, 221–259.
79. Victor Lavy. 2013. Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-based Pay Tournaments among Teachers*. *The Economic Journal* **123**:569, 540–573.
80. Miriam Krieger, Stefan Felder. 2013. Can Decision Biases Improve Insurance Outcomes? An Experiment on Status Quo Bias in Health Insurance Choice. *International Journal of Environmental Research and Public Health* **10**:6, 2560–2577.
81. Linda Nøstbakken. 2013. Formal and informal quota enforcement. *Resource and Energy Economics* **35**:2, 191–215.
82. J. R. Kolstad, I. Lindkvist. 2013. Pro-social preferences and self-selection into the public health sector: evidence from an economic experiment. *Health Policy and Planning* **28**:3, 320–327.
83. Krisztina Kis-Katos, Günther G. Schulze. 2013. Corruption in Southeast Asia: a survey of recent research. *Asian-Pacific Economic Literature* **27**:1, 79–109.

84. Antoine Beretti, Charles Figuières, Gilles Grolleau. 2013. Behavioral innovations: The missing capital in sustainable development?. *Ecological Economics* **89**, 187–195.
85. Daniel F. Stone. 2013. Testing Bayesian Updating With The Associated Press Top 25. *Economic Inquiry* **51**:2, 1457–1474.
86. Holger Stichnoth, Karine Van der Straeten. 2013. Ethnic Diversity, Public Spending, and Individual Support for the Welfare State: A Review of the Empirical Literature. *Journal of Economic Surveys* **27**:2, 364–389.
87. Matthew Doyle, Jacob Wong. 2013. Wage posting without full commitment. *Review of Economic Dynamics* **16**:2, 231–252.
88. J. Kulin, S. Svallfors. 2013. Class, Values, and Attitudes Towards Redistribution: A European Comparison. *European Sociological Review* **29**:2, 155–167.
89. Timo Tammi. 2013. Dictator game giving and norms of redistribution: Does giving in the dictator game parallel with the supporting of income redistribution in the field?. *The Journal of Socio-Economics* **43**, 44–48.
90. Anat Bracha, Chaim Fershtman. 2013. Competitive Incentives: Working Harder or Working Smarter?. *Management Science* **59**:4, 771–781.
91. M. Bhuller, T. Havnes, E. Leuven, M. Mogstad. 2013. Broadband Internet: An Information Superhighway to Sex Crime?. *The Review of Economic Studies*.
92. Arch G. Woodside, Mann Zhang. 2013. Cultural Diversity and Marketing Transactions: Are Market Integration, Large Community Size, and World Religions Necessary for Fairness in Ephemeral Exchanges?. *Psychology & Marketing* **30**:3, 263–276.
93. Cary Deck, Jungmin Lee, Javier A. Reyes, Christopher C. Rosen. 2013. A failed attempt to explain within subject variation in risk taking behavior using domain specific risk attitudes. *Journal of Economic Behavior & Organization* **87**, 1–24.
94. Paul J. Ferraro, Michael K. Price. 2013. Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. *Review of Economics and Statistics* **95**:1, 64–73.
95. Yasuyuki Sawada, Ryuji Kasahara, Keitaro Aoyagi, Masahiro Shoji, Mika Ueyama. 2013. Modes of Collective Action in Village Economies: Evidence from Natural and Artefactual Field Experiments in a Developing Country. *Asian Development Review* **30**:1, 31–51.
96. Dana Chandler, Adam Kapelner. 2013. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets. *Journal of Economic Behavior & Organization*.
97. Thomas S. Dee. 2013. Stereotype Threat and the Student-Athlete. *Economic Inquiry* no-no.
98. Filippos Exadaktylos, Antonio M. Espín, Pablo Brañas-Garza. 2013. Experimental subjects are not different. *Scientific Reports* **3**.
99. Alexander Pfaff, Maria Alejandra Velez, Renzo Taddei, Kenneth Broad. 2013. Unequal Information, Unequal Allocation: Bargaining field experiments in NE Brazil. *Environmental Science & Policy* **26**, 90–101.
100. MoÏse Sagamba, Oleg Shchetinin, Nurmukhammad Yusupov. 2013. Do Microloan Officers Want to Lend to the Less Advantaged? Evidence from a Choice Experiment. *World Development* **42**, 182–198.
101. Katie Baca-Motes, Amber Brown, Ayelet Gneezy, Elizabeth A. Keenan, Leif D. Nelson. 2013. Commitment and Behavior Change: Evidence from the Field. *Journal of Consumer Research* **39**:5, 1070–1084.

102. Peter DeScioli, Siddhi Krishna. 2013. Giving to whom? Altruism in different types of relationships. *Journal of Economic Psychology* **34**, 218–228.
103. Alessandra Righi. 2013. Measuring Social Capital: Official Statistics Initiatives in Italy. *Procedia - Social and Behavioral Sciences* **72**, 4–22.
104. R. Mujcic, P. Frijters. 2013. Economic choices and status: measuring preferences for income rank. *Oxford Economic Papers* **65**:1, 47–73.
105. Ana M. Franco-Watkins, Bryan D. Edwards, Roy E. Acuff. 2013. Effort and Fairness in Bargaining Games. *Journal of Behavioral Decision Making* **26**:1, 79–90.
106. Agnès Festré, Pierre Garrouste. 2013. The Respective Effects of Being Observed and Sanctioned in Modified Dictator and Ultimatum Games. *ISRN Economics* **2013**, 1–13.
107. Johannes Gettinger, Elmar Kiesling, Christian Stummer, Rudolf Vetschera. 2013. A comparison of representations for discrete multi-criteria decision problems. *Decision Support Systems* **54**:2, 976–985.
108. Panayotis Antoniadis, Serge Fdida, Christopher Griffin, Youngmi Jin, George Kesidis. 2013. Distributed medium access control with conditionally altruistic users. *EURASIP Journal on Wireless Communications and Networking* **2013**:1, 202.
109. Ivan Hilliard. 2012. Responsible Management, Incentive Systems, and Productivity. *Journal of Business Ethics*.
110. Kirsten Martin. 2012. Information technology and privacy: conceptual muddles or privacy vacuums?. *Ethics and Information Technology* **14**:4, 267–284.
111. Eric Cardella, Ray Chiu. 2012. Stackelberg in the lab: The effect of group decision making and "Cooling-off " periods. *Journal of Economic Psychology* **33**:6, 1070–1083.
112. Kirsten E. Martin. 2012. Diminished or Just Different? A Factorial Vignette Study of Privacy as a Social Contract. *Journal of Business Ethics* **111**:4, 519–539.
113. Olof Johansson-Stenman, Henrik Svedsäter. 2012. Self-image and valuation of moral goods: Stated versus actual willingness to pay. *Journal of Economic Behavior & Organization* **84**:3, 879–891.
114. Tania Machet, David Lowe, Christian Gütl. 2012. On the potential for using immersive virtual environments to support laboratory experiment contextualisation. *European Journal of Engineering Education* **37**:6, 527–540.
115. Werner Güth, Oliver Kirchkamp. 2012. Will you accept without knowing what? The Yes-No game in the newspaper and in the lab. *Experimental Economics* **15**:4, 656–666.
116. Tanjim Hossain, John A. List. 2012. The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science* **58**:12, 2151–2167.
117. Chaim Fershtman„ Uri Gneezy„ John A. List. 2012. Equity Aversion: Social Norms and the Desire to be Ahead. *American Economic Journal: Microeconomics* **4**:4, 131–144.
118. Thomas Buser. 2012. Digit ratios, the menstrual cycle and social preferences. *Games and Economic Behavior* **76**:2, 457–470.
119. Loukas Balafoutas, Nikos Nikiforakis. 2012. Norm enforcement in the city: A natural field experiment. *European Economic Review* **56**:8, 1773–1785.
120. Gregory A. Huber, Seth J. Hill, Gabriel S. Lenz. 2012. Sources of Bias in Retrospective Decision Making: Experimental Evidence on Voters' Limitations in Controlling Incumbents. *American Political Science Review* **106**:04, 720–741.
121. Leif Brandes, Egon Franck. 2012. Social preferences or personal career concerns? Field evidence on positive and negative reciprocity in the workplace. *Journal of Economic Psychology* **33**:5, 925–939.
122. Hans-Martin Gaudecker, Arthur Soest, Erik Wengström. 2012. Experts in experiments. *Journal of Risk and Uncertainty* **45**:2, 159–190.

123. Guy Grossman, Delia Baldassarri. 2012. The Impact of Elections on Cooperation: Evidence from a Lab-in-the-Field Experiment in Uganda. *American Journal of Political Science* **56**:4, 964–985.
124. Matteo Ploner. 2012. Book Review. *The Journal of Socio-Economics* **41**:5, 693–694.
125. Norbert Hirschauer,, Oliver Mußhoff,. 2012. Smarte Regulierung in der Ernährungswirtschaft durch Name-and-Shame. *Vierteljahrshefte zur Wirtschaftsforschung* **81**:4, 163–181.
126. Lava Yadav, Thomas M. van Rensburg, Hugh Kelley. 2012. A Comparison Between the Conventional Stated Preference Technique and an Inferred Valuation Approach. *Journal of Agricultural Economics* no-no.
127. M. Suchak, F. B. M. de Waal. 2012. Monkeys benefit from reciprocity without the cognitive burden. *Proceedings of the National Academy of Sciences* **109**:38, 15191–15196.
128. Jack Vromen. 2012. Finding the right levers: the serious side of 'economics made fun'. *Journal of Economic Methodology* **19**:3, 199–217.
129. Bruce Lyons, Gordon Douglas Menzies, Daniel John Zizzo. 2012. Conflicting evidence and decisions by agency professionals: an experimental test in the context of merger regulation. *Theory and Decision* **73**:3, 465–499.
130. Tore Ellingsen, Magnus Johannesson, Johanna Mollerstrom, Sara Munkhammar. 2012. Social framing effects: Preferences or beliefs?. *Games and Economic Behavior* **76**:1, 117–130.
131. Lars Schwettmann. 2012. Competing allocation principles: time for compromise?. *Theory and Decision* **73**:3, 357–380.
132. Samuel M. Hartzmark, David H. Solomon. 2012. Efficiency and the Disposition Effect in NFL Prediction Markets. *Quarterly Journal of Finance* **02**:03, 1250013.
133. Eric Cardella. 2012. Learning to make better strategic decisions. *Journal of Economic Behavior & Organization* **84**:1, 382–392.
134. James Alm, Todd L. Cherry, Michael Jones, Michael McKee. 2012. Social programs as positive inducements for tax participation. *Journal of Economic Behavior & Organization* **84**:1, 85–96.
135. Björn Vollan. 2012. Weird reciprocity? A 'within-culture across-country' trust experiment and methodological implications. *Journal of Institutional Economics* **8**:03, 371–397.
136. Mathias Ekström. 2012. Do watching eyes affect charitable giving? Evidence from a field experiment. *Experimental Economics* **15**:3, 530–546.
137. Souleiman Naciri, Min-Jung Yoo, Rémy GlardonUsing Serious Games for Collecting and Modeling Human Procurement Decisions in a Supply Chain Context 744–765.
138. Gary Charness,, Matthias Sutter. 2012. Groups Make Better Self-Interested Decisions. *Journal of Economic Perspectives* **26**:3, 157–176.
139. Xiaobo Lü, Kenneth Scheve, Matthew J. Slaughter. 2012. Inequity Aversion and the International Distribution of Trade Protection. *American Journal of Political Science* **56**:3, 638–654.
140. Samuel Bowles,, Sandra Polanía-Reyes. 2012. Economic Incentives and Social Preferences: Substitutes or Complements?. *Journal of Economic Literature* **50**:2, 368–425.
141. Yannick Ferreira De Sousa, Alistair Munro. 2012. Truck, barter and exchange versus the endowment effect: Virtual field experiments in an online game environment. *Journal of Economic Psychology* **33**:3, 482–493.
142. Devah Pager, Bruce Western. 2012. Identifying Discrimination at Work: The Use of Field Experiments. *Journal of Social Issues* **68**:2, 221–237.

143. Donald V. Moser, Patrick R. Martin. 2012. A Broader Perspective on Corporate Social Responsibility Research in Accounting. *The Accounting Review* **87**:3, 797–806.
144. Franziska Barmettler, Ernst Fehr, Christian Zehnder. 2012. Big experimenter is watching you! Anonymity and prosocial behavior in the laboratory. *Games and Economic Behavior* **75**:1, 17–34.
145. Steffen Huck, Wieland Müller. 2012. Allais for all: Revisiting the paradox in a large representative sample. *Journal of Risk and Uncertainty*.
146. References 577–616.
147. Aaron Nicholas. 2012. Fairness as a constraint on reciprocity: Playing simultaneously as dictator and trustee. *The Journal of Socio-Economics* **41**:2, 211–221.
148. B.P. Soh, W. Lee, P.L. Kench, W.M. Reed, M.F. McEntee, A. Poulos, P.C. Brennan. 2012. Assessing reader performance in radiology, an imperfect science: Lessons from breast screening. *Clinical Radiology*.
149. Bart J. Wilson. 2012. Contra Private Fairness*. *American Journal of Economics and Sociology* **71**:2, 407–435.
150. Shaun Hargreaves Heap, Arjan Verschoor, Daniel John Zizzo. 2012. A test of the experimental method in the spirit of Popper. *Journal of Economic Methodology* **19**:1, 63–76.
151. Souleiman Naciri, Min-Jung Yoo, Rémy GlardonUsing Serious Games for Collecting and Modeling Human Procurement Decisions in a Supply Chain Context 135–156.
152. David Gill„ Victoria Prowse. 2012. A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review* **102**:1, 469–503.
153. Joachim I. Krueger, Theresa E. DiDonato, David Freestone. 2012. Social Projection Can Solve Social Dilemmas. *Psychological Inquiry* **23**:1, 1–27.
154. Robert Hoffmann. 2012. The Experimental Economics of Religion. *Journal of Economic Surveys* no-no.
155. Olivier Armantier, Amadou BolyChapter 5 On the External Validity of Laboratory Experiments on Corruption **15**, 117–144.
156. Ananish ChaudhuriChapter 2 Gender and Corruption: A Survey of the Experimental Evidence **15**, 13–49.
157. Claudia Toma, Karl-Andrew Woltin. 2012. Motivational and Contextual Considerations Concerning the Social Projection Hypothesis. *Psychological Inquiry* **23**:1, 69–74.
158. Kevin Holmes, Lisa Marriott, John Randal. 2012. Ethics and experiments in accounting: A contribution to the debate on measuring ethical behaviour. *Pacific Accounting Review* **24**:1, 80–100.
159. D. H. Herberich, J. A. List. 2012. Digging into Background Risk: Experiments with Farmers and Students. *American Journal of Agricultural Economics* **94**:2, 457–463.
160. Pieter A. Gautier, Bas van der Klaauw. 2012. Selection in a field experiment with voluntary participation. *Journal of Applied Econometrics* **27**:1, 63–84.
161. Martijn J. van den Assem, Dennie van Dolder, Richard H. Thaler. 2012. Split or Steal? Cooperative Behavior When the Stakes Are Large. *Management Science* **58**:1, 2–20.
162. David Cesarini, Magnus Johannesson, Patrik K. E. Magnusson, Björn Wallace. 2012. The Behavioral Genetics of Behavioral Anomalies. *Management Science* **58**:1, 21–34.
163. Ayelet Gneezy, Alex Imas, Amber Brown, Leif D. Nelson, Michael I. Norton. 2012. Paying to Be Nice: Consistency and Costly Prosocial Behavior. *Management Science* **58**:1, 179–187.

164. Cecile Jackson. 2011. Internal and External Validity in Experimental Games: A Social Reality Check. *European Journal of Development Research*.
165. Steffen Andersen„ Seda Ertaç„ Uri Gneezy„ Moshe Hoffman„ John A. List. 2011. Stakes Matter in Ultimatum Games. *American Economic Review* **101**:7, 3427–3439.
166. Nisvan Erkal„ Lata Gangadharan„ Nikos Nikiforakis. 2011. Relative Earnings and Giving in a Real- Effort Experiment. *American Economic Review* **101**:7, 3330–3348.
167. Kewen Wu, Yuxiang Zhao, Qinghua Zhu, Xiaojie Tan, Hua Zheng. 2011. A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management* **31**:6, 572–581.
168. Quang Nguyen. 2011. Does nurture matter: Theory and experimental investigation on the effect of working environment on risk and time preferences. *Journal of Risk and Uncertainty*.
169. John Kerr, Mamta Vardhan, Rohit Jindal. 2011. Prosocial behavior and incentives: Evidence from field experiments in rural Mexico and Tanzania. *Ecological Economics*.
170. Jeff Galak, Deborah Small, Andrew T Stephen. 2011. Microfinance Decision Making: A Field Study of Prosocial Lending. *Journal of Marketing Research* **48**:SPL, S130-S137.
171. P. Edwards. 2011. Experimental economics and workplace behaviour: bridges over troubled methodological waters?. *Socio-Economic Review*.
172. Brian T Kench, Neil B Niman. 2011. Of Altruists and Thieves. *Eastern Economic Journal* **36**:3, 317–343.
173. Dan Kirk, Peter M. Gollwitzer, Peter J. Carnevale. 2011. Self-Regulation in Ultimatum Bargaining: Goals and Plans Help Accepting Unfair but Profitable Offers. *Social Cognition* **29**:5, 528–546.
174. Maarten Voors, Ty Turley, Andreas Kontoleon, Erwin Bulte, John A. List. 2011. Behavior in context- free experiments is not predictive of behavior in the field: Evidence from public good experiments in rural Sierra Leone. *Economics Letters*.
175. Roger Berger, Heiko Rauhut, Sandra Prade, Dirk Helbing. 2011. Bargaining over waiting time in ultimatum game experiments. *Social Science Research*.
176. Hunt Allcott. 2011. Social norms and energy conservation. *Journal of Public Economics* **95**:9–10, 1082–1095.
177. Cynthia E. Cryder, George Loewenstein. 2011. Responsibility: The tie that binds. *Journal of Experimental Social Psychology*.
178. Loren King. 2011. Exploitation and Rational Choice. *Canadian Journal of Political Science* **44**:03, 635–661.
179. María Jiménez-Buedo. 2011. Conceptual tools for assessing experiments: some well-entrenched confusions regarding the internal/external validity distinction. *Journal of Economic Methodology* **18**:3, 271–282.
180. Veronika A. Andorfer, Ulf Liebe. 2011. Research on Fair Trade Consumption—A Review. *Journal of Business Ethics*.
181. Robert Östling„ Joseph Tao-yi Wang„ Eileen Y. Chou„ Colin F. Camerer. 2011. Testing Game Theory in the Field: Swedish LUPI Lottery Games. *American Economic Journal: Microeconomics* **3**:3, 1–33.
182. M. Shayo, A. Zussman. 2011. Judicial Ingroup Bias in the Shadow of Terrorism. *The Quarterly Journal of Economics* **126**:3, 1447–1484.
183. A. Tavoni, A. Dannenberg, G. Kallis, A. Loschel. 2011. From the Cover: Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences* **108**:29, 11825–11829.

184. D. Baldassarri, G. Grossman. 2011. Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences* **108**:27, 11023–11027.
185. T.K. Ahn, Elinor Ostrom, James Walker. 2011. Reprint of: A common-pool resource experiment with postgraduate subjects from 41 countries. *Ecological Economics* **70**:9, 1580–1589.
186. John Duffy. 2011. Trust in Second Life. *Southern Economic Journal* **78**:1, 53–62.
187. Pavel Atanasov, Jason Dana. 2011. Leveling the Playing Field: Dishonesty in the Face of Threat. *Journal of Economic Psychology*.
188. Daniel Castillo, François Bousquet, Marco A. Janssen, Kobchai Worrapimphong, Juan Camillo Cardenas. 2011. Context matters to explain field experiments: Results from Colombian and Thai fishing villages. *Ecological Economics* **70**:9, 1609–1620.
189. Vera Angelova, Werner Güth, Martin G. Kocher. 2011. Co-employment of permanently and temporarily employed agents. *Labour Economics*.
190. K. L. Milkman, J. Beshears, J. J. Choi, D. Laibson, B. C. Madrian. 2011. Using implementation intentions prompts to enhance influenza vaccination rates. *Proceedings of the National Academy of Sciences* **108**:26, 10415–10420.
191. Niclas Berggren, Christian Bjørnskov. 2011. Is the importance of religion in daily life related to social trust? Cross-country and cross-state comparisons. *Journal of Economic Behavior & Organization*.
192. Maik Dierkes, Alexander Klos, Thomas Langer. 2011. A note on representativeness and household finance. *Economics Letters*.
193. Noel D. Johnson, Alexandra A. Mislin. 2011. Trust Games: A Meta-Analysis. *Journal of Economic Psychology*.
194. Glenn W. Harrison. 2011. The methodological promise of experimental economics. *Journal of Economic Methodology* **18**:2, 183–187.
195. Stephanie H. Fay, Graham Finlayson. 2011. Negative affect-induced food intake in non-dieting women is reward driven and associated with restrained–disinhibited eating subtype. *Appetite* **56**:3, 682–688.
196. Christoph Engel. 2011. Dictator games: a meta study. *Experimental Economics*.
197. Maarten Voors„ Erwin Bulte„ Andreas Kontoleon„ John A. List„ Ty Turley. 2011. Using Artefactual Field Experiments to Learn about the Incentives for Sustainable Forest Use in Developing Economies. *American Economic Review* **101**:3, 329–333.
198. Omar Al-Ubaydli„ Min Lee. 2011. Can Tailored Communications Motivate Environmental Volunteers? A Natural Field Experiment. *American Economic Review* **101**:3, 323–328.
199. Pamela Jakiela. 2011. Social Preferences and Fairness Norms as Informal Institutions: Experimental Evidence. *American Economic Review* **101**:3, 509–513.
200. Olivier Armantier, Amadou Boly. 2011. A controlled field experiment on corruption. *European Economic Review*.
201. John A. List, Charles F. Mason. 2011. Are CEOs expected utility maximizers?. *Journal of Econometrics* **162**:1, 114–123.
202. James Alm. 2011. Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies. *International Tax and Public Finance*.
203. Nick Feltovich. 2011. What'S to Know About Laboratory Experimentation in Economics?. *Journal of Economic Surveys* **25**:2, 371–379.
204. Johann Graf Lambsdorff, Björn Frank. 2011. Corrupt reciprocity – Experimental evidence on a men's game. *International Review of Law and Economics*.

205. María Valle Santos Álvarez, María Teresa García Merino, Eleuterio Vallelado González. 2011. La percepción directiva: influencia del perfil cognitivo y de factores contextuales. *Cuadernos de Economía y Dirección de la Empresa* **14**:2, 67–77.
206. John A. List, Sally Sadoff, Mathis Wagner. 2011. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*.
207. Joseph Eisenhauer, Doris Geide-Stevenson, David Ferro. 2011. Experimental Estimates of Taxpayer Ethics. *Review of Social Economy* **69**:1, 29–53.
208. Floris Heukelom. 2011. How validity travelled to economic experimenting. *Journal of Economic Methodology* **18**:1, 13–28.
209. Sally Gainsbury, Alex Blaszczynski. 2011. The Appropriateness of Using Laboratories and Student Participants in Gambling Research. *Journal of Gambling Studies* **27**:1, 83–97.
210. Jayson L. Lusk, Brian C. Briggeman. 2011. Selfishness, altruism, and inequality aversion toward consumers and farmers. *Agricultural Economics* **42**:2, 121–139.
211. Henk Folmer, Olof Johansson-Stenman. 2011. Does Environmental Economics Produce Aeroplanes Without Engines? On the Need for an Environmental Social Science. *Environmental and Resource Economics* **48**:3, 337–361.
212. Ben Greiner, Axel Ockenfels, Peter Werner. 2011. Wage Transparency and Performance: A Real- Effort Experiment. *Economics Letters*.
213. Björn Tyrefors Hinnerich, Erik Höglin, Magnus Johannesson. 2011. Are boys discriminated in Swedish high schools?. *Economics of Education Review*.
214. Pallab Mozumder, Robert Berrens. 2011. Social context, financial stakes and hypothetical bias: an induced value referendum experiment. *Applied Economics* 1–13.
215. James Alm, Benno Torgler. 2011. Do Ethics Matter? Tax Compliance and Morality. *Journal of Business Ethics*.
216. Phillipa Caudwell, Catherine Gibbons, Mark Hopkins, Erik Naslund, Neil King, Graham Finlayson, John Blundell. 2011. The influence of physical activity on appetite control: an experimental system to understand the relationship between exercise-induced energy expenditure and energy intake. *Proceedings of the Nutrition Society* 1–10.
217. Mathilde Almlund, Angela Lee Duckworth, James Heckman, Tim KautzPersonality Psychology and Economics **4**, 1–181.
218. John A. List, Imran RasulField Experiments in Labor Economics **4**, 103–228.
219. Jan-Erik Lönnqvist, Markku Verkasalo, Gari Walkowitz. 2011. It pays to pay – Big Five personality influences on co-operative behaviour in an incentivized and hypothetical prisoner's dilemma gameĄŹ. *Personality and Individual Differences* **50**:2, 300–304.
220. Gary Charness, Peter KuhnLab Labor: What Can Labor Economists Learn from the Lab? **4**, 229–330.
221. Luigi Guiso, Paola Sapienza, Luigi ZingalesCivic Capital as the Missing Link **1**, 417–480.
222. Reviva Hasson, Åsa Löfgren, Martine Visser. 2010. Climate change in a public goods game: Investment decision in mitigation versus adaptationĄŹ. *Ecological Economics* **70**:2, 331–338.
223. Nava Ashraf,, James Berry,, Jesse M. Shapiro. 2010. Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia. *American Economic Review* **100**:5, 2383–2413.

224. Tomas Dvorak, Henry Hanley. 2010. Financial literacy and the design of retirement plans. *The Journal of Socio-Economics* **39**:6, 645–652.
225. Maria Claudia Lopez, James J. Murphy, John M. Spraggon, John K. Stranlund. 2010. Comparing the Effectiveness of Regulation and Pro- Social Emotions To Enhance Cooperation: Experimental Evidence From Fishing Communities in Colombia. *Economic Inquiry* no-no.
226. Dean Karlan, John A. List, Eldar Shafir. 2010. Small matches and charitable giving: Evidence from a natural field experimentĄŹ. *Journal of Public Economics*.
227. Amadou Boly. 2010. On the incentive effects of monitoring: evidence from the lab and the field. *Experimental Economics*.
228. Stuart A. West, Claire El Mouden, Andy Gardner. 2010. Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*.
229. Andrew F. Reeson, John G. Tisdell. 2010. The Market Instinct: The Demise of Social Preferences for Self-Interest. *Environmental and Resource Economics* **47**:3, 439–453.
230. Kenneth L. Leonard, Melkiory C. Masatu. 2010. Using the Hawthorne effect to examine the gap between a doctor's best possible practice and actual performance. *Journal of Development Economics* **93**:2, 226–234.
231. Nicola Lacetera, Mario Macis. 2010. Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization* **76**:2, 225–237.
232. Fredrik Carlsson, Jorge H. García, Åsa Löfgren. 2010. Conformity and the Demand for Environmental Goods. *Environmental and Resource Economics* **47**:3, 407–421.
233. T.K. Ahn, Elinor Ostrom, James Walker. 2010. A common-pool resource experiment with postgraduate subjects from 41 countries. *Ecological Economics* **69**:12, 2624–2633.
234. David Cesarini, Magnus Johannesson, Paul Lichtenstein, ÖRjan Sandewall, BjÖRn Wallace. 2010. Genetic Variation in Financial Decision-Making. *The Journal of Finance* **65**:5, 1725–1754.
235. Marco Castillo, Gregory Leo. 2010. Moral Hazard and Reciprocity. *Southern Economic Journal* **77**:2, 271–281.
236. Soosung Hwang, Steve E. Satchell. 2010. How loss averse are investors in financial markets?. *Journal of Banking & Finance* **34**:10, 2425–2438.
237. Richard E. Just. 2010. Behavior, Robustness, and Sufficient Statistics in Welfare Measurement. *Annual Review of Resource Economics* **3**:1, 110301102409075.
238. Mitchell Hoffman. 2010. Does Higher Income Make You More Altruistic? Evidence from the Holocaust. *Review of Economics and Statistics* 110510162004036.
239. Gabriella Conti, Stephen Pudney. 2010. Survey Design and the Analysis of Satisfaction. *Review of Economics and Statistics* 110510162004036.
240. Gabriel Katz, R. Michael Alvarez, Ernesto Calvo, Marcelo Escolar, Julia Pomares. 2010. Assessing the Impact of Alternative Voting Technologies on Multi-Party Elections: Design Features, Heuristic Processing and Voter Choice. *Political Behavior*.
241. Xavier Giné, Pamela Jakiela, Dean Karlan, Jonathan Morduch. 2010. Microfinance Games. *American Economic Journal: Applied Economics* **2**:3, 60–95.
242. Pablo Brañas-Garza, Teresa García-Muñoz, Shoshana Neuman. 2010. The big carrot: High-stakes incentives revisited. *Journal of Behavioral Decision Making* **23**:3, 288–313.

243. Craig E. Landry, Andreas Lange, John A. List, Michael K. Price, Nicholas G. Rupp. 2010. Is a Donor in Hand Better than Two in the Bush? Evidence from a Natural Field Experiment. *American Economic Review* **100**:3, 958–983.
244. Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde. 2010. Are Risk Aversion and Impatience Related to Cognitive Ability?. *American Economic Review* **100**:3, 1238–1260.
245. Jack Vromen. 2010. Where economics and neuroscience might meet. *Journal of Economic Methodology* **17**:2, 171–183.
246. Moana Vercoe, Paul Zak. 2010. Inductive modeling using causal studies in neuroeconomics: brains on drugs. *Journal of Economic Methodology* **17**:2, 133–146.
247. Cameron Hepburn, Stephen Duncan, Antonis Papachristodoulou. 2010. Behavioural Economics, Hyperbolic Discounting and Environmental Policy. *Environmental and Resource Economics* **46**:2, 189–206.
248. Fredrik Carlsson. 2010. Design of Stated Preference Surveys: Is There More to Learn from Behavioral Economics?. *Environmental and Resource Economics* **46**:2, 167–177.
249. Joseph Henrich, Steven J. Heine, Ara Norenzayan. 2010. Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences* **33**:2–3, 111–135.
250. R. Kummerli, M. N. Burton-Chellew, A. Ross-Gillespie, S. A. West. 2010. Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games. *Proceedings of the National Academy of Sciences* **107**:22, 10125–10130.
251. Manuel Souto-Otero. 2010. Education, meritocracy and redistribution. *Journal of Education Policy* **25**:3, 397–413.
252. Jason Barabas, Jennifer Jerit. 2010. Are Survey Experiments Externally Valid?. *American Political Science Review* **104**:02, 226–242.
253. James Konow. 2010. Mixed feelings: Theories of and evidence on giving. *Journal of Public Economics* **94**:3–4, 279–297.
254. Stacey L. Rucas, Michael Gurven, Hillard Kaplan, Jeffrey Winking. 2010. The Social Strategy Game. *Human Nature* **21**:1, 1–18.
255. Kirsten Bregn. 2010. The Logic of the New Pay Systems Revisited-in the Light of Experimental and Behavioral Economics. *International Journal of Public Administration* **33**:4, 161–168.
256. Daniel John Zizzo. 2010. Experimenter demand effects in economic experiments. *Experimental Economics* **13**:1, 75–98.
257. Jacob K. Goeree„ Margaret A. McConnell„ Tiffany Mitchell„ Tracey Tromp„ Leeat Yariv. 2010. The 1/d Law of Giving. *American Economic Journal: Microeconomics* **2**:1, 183–203.
258. Carina Cavalcanti, Felix Schläpfer, Bernhard Schmid. 2010. Public participation and willingness to cooperate in common-pool resource management: A field experiment with fishing communities in Brazil. *Ecological Economics* **69**:3, 613–622.
259. Céline Michaud, Daniel Llerena. 2010. Green consumer behaviour: an experimental analysis of willingness to pay for remanufactured products. *Business Strategy and the Environment* n/a-n/a.
260. Maroš Servátka. 2010. Does generosity generate generosity? An experimental study of reputation effects in a dictator game. *The Journal of Socio-Economics* **39**:1, 11–17.
261. Fredrik Carlsson, Dinky Daruvala, Henrik Jaldell. 2010. Do you do what you say or do you do what you say others do?. *Journal of Choice Modelling* **3**:2, 113–133.

262. Anthony M. Evans, Joachim I. Krueger. 2009. The Psychology (and Economics) of Trust. *Social and Personality Psychology Compass* **3**:6, 1003–1017.
263. Cecile Jackson. 2009. Researching the Researched: Gender, Reflexivity and Actor-Orientation in an Experimental Game. *European Journal of Development Research* **21**:5, 772–791.
264. Stephen Toler, Brian C. Briggeman, Jayson L. Lusk, Damian C. Adams. 2009. Fairness, Farmers Markets, and Local Production. *American Journal of Agricultural Economics* **91**:5, 1272–1278.
265. David R. Just, Steven Y. Wu. 2009. Experimental Economics and the Economics of Contracts. *American Journal of Agricultural Economics* **91**:5, 1382–1388.
266. Rockoff Jonah. 2009. Field Experiments in Class Size from the Early Twentieth Century. *Journal of Economic Perspectives* **23**:4, 211–230.
267. Mahmud Yesuf, Randall A. Bluffstone. 2009. Poverty, Risk Aversion, and Path Dependence in Low- Income Countries: Experimental Evidence from Ethiopia. *American Journal of Agricultural Economics* **91**:4, 1022–1037.
268. Agns Festr. 2009. Incentives and Social Norms: A Motivation-Based Economic Analysis of Social Norms. *Journal of Economic Surveys*.
269. A. Falk, J. J. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* **326**:5952, 535–538.
270. Juan Camilo Cárdenas. 2009. Experiments in Environment and Development. *Annual Review of Resource Economics* **1**:1, 157–182.
271. Christopher Timmins, Wolfram Schlenker. 2009. Reduced-Form Versus Structural Modeling in Environmental and Resource Economics. *Annual Review of Resource Economics* **1**:1, 351–380.
272. Jen Shang, Rachel Croson. 2009. A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods. *The Economic Journal* **119**:540, 1422–1439.
273. John Ermisch, Diego Gambetta, Heather Laurie, Thomas Siedler, S. C. Noah Uhrig. 2009. Measuring people's trust. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**:4, 749–769.
274. John A. List. 2009. Social Preferences: Some Thoughts from the Field. *Annual Review of Economics* **1**:1, 563–579.
275. Ernst Fehr, Lorenz Goette, Christian Zehnder. 2009. A Behavioral Account of the Labor Market: The Role of Fairness Concerns. *Annual Review of Economics* **1**:1, 355–384.
276. Ori Heffetz„ Moses Shayo. 2009. How Large Are Non-Budget-Constraint Effects of Prices on Demand?. *American Economic Journal: Applied Economics* **1**:4, 170–199.
277. Jayson L. Lusk, F. Bailey Norwood. 2009. Bridging the gap between laboratory experiments and naturally occurring markets: An inferred valuation method. *Journal of Environmental Economics and Management* **58**:2, 236–250.
278. Jeffrey C. Ely„ Tanjim Hossain. 2009. Sniping and Squatting in Auction Markets. *American Economic Journal: Microeconomics* **1**:2, 68–94.
279. David Cesarini, Christopher T. Dawes, Magnus Johannesson, Paul Lichtenstein, Björn Wallace. 2009. Experimental Game Theory and Behavior Genetics. *Annals of the New York Academy of Science* **1167**:1, 66–75.
280. Rachel Croson„ Uri Gneezy. 2009. Gender Differences in Preferences. *Journal of Economic Literature* **47**:2, 448–474.

281. Stefano DellaVigna. 2009. Psychology and Economics: Evidence from the Field. *Journal of Economic Literature* **47**:2, 315–372.
282. Jae Bong Chang, Jayson L. Lusk, F. Bailey Norwood. 2009. How Closely Do Hypothetical Surveys and Laboratory Experiments Predict Field Behavior?. *American Journal of Agricultural Economics* **91**:2, 518–534.
283. B. Kelsey Jack. 2009. Upstream–downstream transactions and watershed externalities: Experimental evidence from KenyaĄŹ. *Ecological Economics* **68**:6, 1813–1824.
284. S.-W. Wu, M. R. Delgado, L. T. Maloney. 2009. Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences* **106**:15, 6088–6093.
285. James Sundali, Federico Guerrero. 2009. Managing a 401(k) Account: An Experiment on Asset Allocation. *Journal of Behavioral Finance* **10**:2, 108–124.
286. Christina M. Fong„ Erzo F. P. Luttmer. 2009. What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty. *American Economic Journal: Applied Economics* **1**:2, 64–87.
287. Abigail Barr, Pieter Serneels. 2009. Reciprocity in the workplace. *Experimental Economics* **12**:1, 99–112.
288. A Dreber, C Apicella, D Eisenberg, J Garcia, R Zamore, J Lum, B Campbell. 2009. The 7R polymorphism in the dopamine receptor D4 gene (DRD4) is associated with financial risk taking in menĄŹ. *Evolution and Human Behavior* **30**:2, 85–92.
289. Dmitry A. Shapiro. 2009. The role of utility interdependence in public good experiments. *International Journal of Game Theory* **38**:1, 81–106.
290. Katherine Silz Carson, Susan M. Chilton, W. George Hutchinson. 2009. Necessary conditions for demand revelation in double referenda. *Journal of Environmental Economics and Management* **57**:2, 219–225.
291. Kate Antonovics, Peter Arcidiacono, Randall Walsh. 2009. The Effects of Gender Interactions in the Lab and in the Field. *Review of Economics and Statistics* **91**:1, 152–162.
292. Jim Engle-Warnick, Javier Escobal, Sonia Laszlo. 2009. How do additional alternatives affect individual choice under uncertainty?. *Canadian Journal of Economics/Revue canadienne d'économique* **42**:1, 113–140.
293. Marc Ohana. 2009. La réciprocité sur le marché du travail : les limites du laboratoire. *L'Actualité économique* **85**:2, 239.
294. Bernard J. (Jim) Jansen. 2009. Understanding User-Web Interactions via Web Analytics. *Synthesis Lectures on Information Concepts, Retrieval, and Services* **1**:1, 1–102.
295. Stephen M. Fiore, Glenn W. Harrison, Charles E. Hughes, E. Elisabet Rutström. 2009. Virtual experiments and environmental policyĄŹ. *Journal of Environmental Economics and Management* **57**:1, 65–86.
296. Richard E. Just. 2008. Distinguishing Preferences from Perceptions for Meaningful Policy Analysis. *American Journal of Agricultural Economics* **90**:5, 1165–1175.
297. Andreas Ortmann. 2008. Prospecting Neuroeconomics. *Economics and Philosophy* **24**:03, 431.
298. J Ladenburg, S Olsen. 2008. Gender-specific starting point bias in choice experiments: Evidence from an empirical study. *Journal of Environmental Economics and Management* **56**:3, 275–285.

299. Walter Borges, Harold Clarke. 2008. Cues in Context: Analyzing the Heuristics of Referendum Voting with an Internet Survey Experiment. *Journal of Elections, Public Opinion & Parties* **18**:4, 433–448.
300. G. Moschini. 2008. Biotechnology and the development of food markets: retrospect and prospects. *European Review of Agricultural Economics* **35**:3, 331–355.
301. Rachel Croson, Jen (Yue) Shang. 2008. The impact of downward social information on contribution decisions. *Experimental Economics* **11**:3, 221–233.
302. J Bouma, E Bulte, D Vansoest. 2008. Trust and cooperation: Social capital and community resource management. *Journal of Environmental Economics and Management* **56**:2, 155–166.
303. Matthias Benz, Stephan Meier. 2008. Do people behave in experiments as in the field?—evidence from donations. *Experimental Economics* **11**:3, 268–281.
304. John A. List. 2008. Introduction to field experiments in economics with applications to the economics of charity. *Experimental Economics* **11**:3, 203–212.
305. Jeffrey Carpenter, Cristina Connolly, Caitlin Knowles Myers. 2008. Altruistic behavior in a representative dictator experiment. *Experimental Economics* **11**:3, 282–298.
306. Daniel Rondeau, John A. List. 2008. Matching and challenge gifts to charity: evidence from laboratory and natural field experiments. *Experimental Economics* **11**:3, 253–267.
307. Daijiro Okada, Paul M. Bingham. 2008. Human uniqueness-self-interest and social cooperation. *Journal of Theoretical Biology* **253**:2, 261–270.
308. Roland G. Fryer„ Steven D. Levitt„ John A. List. 2008. Exploring the Impact of Financial Incentives on Stereotype Threat: Evidence from a Pilot Study. *American Economic Review* **98**:2, 370–375.
309. K Leonard. 2008. Is patient satisfaction sensitive to changes in the quality of care? An exploitation of the Hawthorne effect#. *Journal of Health Economics* **27**:2, 444–459.
310. Steffen Andersen, Glenn W. Harrison, Morten I. Lau, E. Elisabet RutströmRisk aversion in game shows **12**, 359–404.
311. Dean Karlan, John A. List. 2007. Does Price Matter in Charitable Giving? Evidence from a Large- Scale Natural Field Experiment. *American Economic Review* **97**:5, 1774–1793.

CHAPTER 14

THE PROMISE AND SUCCESS OF LAB–FIELD GENERALIZABILITY IN EXPERIMENTAL ECONOMICS: A CRITICAL REPLY TO LEVITT AND LIST

COLIN F. CAMERER

Introduction

Lab experimentation took serious hold in economics much later than in its neighboring social sciences and in biology. This late development was associated with slow recognition of the possibility and nature of economic experiments. Only about 20 years ago, the popular economics text by Samuelson and Nordhaus (1985, p. 8) stated:

> One possible way of figuring out economic laws . . . is by controlled experiments . . . Economists [unfortunately] . . . cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe.

When it was published in 1985, this passage would have surprised and disappointed the hundreds of economists who actually *had already* done controlled experiments and had published their results in leading journals.

The belated development of experimental economics also seems to be associated with a persistent confusion about the styles of psychology and economics experiments. The blurring of this methodological boundary between psychology and economics may account for some of the ongoing skepticism about generalizability of economics experiments. For example, Gary Becker was quoted in a casual magazine interview as saying:

> One can get excellent suggestions from experiments, but economic theory is not about how people act in experiments, but how they act in markets. And those are very different things. It is similar to asking people why they do things. That may be useful to get suggestions, but it is not a test of the theory. The theory is not about how people answer questions. It is a theory about how people actually choose in market situations. (Clement, 2002)

Becker says that "how people act in experiments" is similar to "asking people why they do things." Lumping the two types of data together is distressing to experimental economics because our approach is dedicated precisely to creating features of experimental economics which are distinct from surveys (and from many features of experiments in psychology). Indeed, the great methodological achievement of experimental economics has been the creation of experimental markets which are designed to have the key features—institutional rules, endowments, and incentives—that are considered essential to make predictions from economic theory about behavior (e.g., Smith, 1982).

This chapter addresses a recent, vocal criticism about experimental economics: Lab experimental findings may not generalize to field settings. Levitt and List (2007a) expressed this view in an especially provocative and ominous way:

> Yet unless considerable changes are made in the manner in which we conduct lab experiments, our model highlights that the relevant factors will rarely converge across the lab and many field settings. (Levitt and List, 2007a, p. 364)

Interestingly, a roughly similar debate about generalizability is ongoing in development economics. The debate is about the value of randomized field experiments compared to inferences from larger observational samples. In this debate the imperfections of field experiments are sometimes shown in a harsh light, rather than as a flattering contrast to lab experiments. For example, in a news interview Angus Deaton said the following:

> They [field experiments in developing countries] tend to be very good at proving causality but often in a very narrow framework. Just because something works in this crazy, local way doesn't necessarily mean that is very useful. (Stanley, 2011)

Defending these experiments, Banerjee and Duflo suggest that

> . . . because of an imperfect recognition of what is exciting about the experimental agenda, there is a tendency to set up false oppositions between experimental work and other forms of research. (Banerjee and Duflo, 2009, p. 152)

This chapter has three main points:

First, special concern about generalizability of lab results might result from an aversion to the stylized perspective on what economics experiments are meant to do, which most experimentalists hold (which I call the scientific view).

The *policy view* is that generalizability is crucial, because an implicit goal of lab experiments is to extrapolate from the lab to a particular field setting (or to some imagined setting which is the target of "external validity"). 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 behavior. This general behavior function is assumed to be parallel in the lab and field. In this view, since the goal is to understand general principles, whether the "lab generalizes to the field" (sometimes called "external validity" of an experiment) is distracting, difficult to know (since there is no single "external" target field setting), and is no more useful than asking whether "the field generalizes to the lab."

Second, even in the scientific perspective, it is certainly true that *some* economics experiments have features that make them less likely to generalize to *some* naturally occurring settings. However, the exact same thing is true of *some* field settings that will not generalize well to *some* field settings.

We then consider which common features of lab experiments might threaten generalizability. Are those features a necessary part of all lab experiments? Except for obtrusive observation in the lab—which is an inherent result of federally regulated human subjects protection in the United States (though not in all countries)—the answer is "no." The special features of lab experiments which might limit generalizability can therefore be relaxed, if necessary, to more closely match particular field settings. Then we ask whether typical lab features *necessarily* undermine generalizability to all field settings, in a way that cannot be corrected. They do not.

Third, we review economics experiments which are specifically designed to generalize from the lab to the field. Only a few experiments were deliberately designed to have properties of field settings and were conducted in parallel naturally occurring field environments. List's (2006) study of reciprocal quality delivery in sports card and ticket markets is looked at in detail. This extra scrutiny is justified because this is the earliest carefully designed study that compares lab and field differences feature-by-feature, was prominently published in the Journal of Political

Economy, and plays a large evidential role in Levitt and List (2007a,b, 2008). In that study, there is no general lab–field difference in reciprocity for all card dealers. There *is* a difference for a minority sample of nonlocal dealers, but it is statistically unreliable (based on new analyses not previously reported, which are within subjects and hence have good power). Therefore that study does *not* show conclusive evidence of poor lab generalizability. A meticulous study of Dutch fishermen (Stoop et al., 2010) does show more prosociality in the lab than in a comparable field setting. That is the *only* study which shows reliable evidence of poor generalization when lab and field features are closely matched.

Furthermore, there are many experiments comparing lab behavior with one group of subjects with similar field behavior from different sample of participants, but with no effort to match the lab and field closely. In this type of comparison, there are more than 20 examples of good comparability (particularly for social preference expression) and only two examples with poor comparability.

Finally, note that the phrase "promise of generalizability" in my title describes two kinds of promise. The first "promise" is whether every lab experiment actually promises to generalize to a particular field setting. The answer is "no": The typical experiment promises only to deliver interesting data about the general mapping from variables to behavior, with extraordinary control over independent variables, not to generalize to any particular target setting. The second kind of "promise" is whether generalizability is likely to be "promising" (i.e., typically accurate) for those lab experiments that are specifically designed to have features much like those in closely parallel settings. The answer is "yes."

Is Generalizability Fundamental? The Scientific View

Experimental economists have converged on a reasonably consensual view of what role experimental data play in economics, which minimizes the importance of narrow worries about external validity. In this view, all empirical studies are designed to contribute evidence about the *general* way in which individual characteristics, incentives, rules, and endowments influence economic behavior. Experiments contribute especially diagnostic evidence by virtue of extraordinary control over independent variables (internal validity). The price of this control is a necessary sacrifice in obtrusive measurement. There is an offsetting bonus of near-perfect replicability.

The guiding idea here is what Vernon Smith (1976, 1982) called "parallelism": It is the hopeful assumption that the same general laws apply in all settings. To be sure, the "general laws" can certainly include how behavior depends on a wide variety of parameter values that often covary with lab and field features, but are potentially controllable or measurable as nuisance variables. For example, parallelism does *not* require that students in a lab setting designed to resemble

foreign exchange traders behave in the same way as professional foreign exchange traders behave on trading floors. Behavior may differ because the subject populations are different, or because of many differences in the lab and the trading floor. The maintained assumption of parallelism simply asserts that if those differences could be held constant (or controlled for econometrically), behavior in the lab and the trading floor would be the same. Put differently, if many experimental and field data sets were combined, with sufficient variation among variables like stakes, experience, and subject characteristics, a "Lab" dummy variable would not be significant (assuming that it did not correlate with omitted variables; Al-Ubaydli and List (2015) call this "investigation-neutrality"). In this very specific econometric sense, the maintained hypothesis in experimental economics is that there is absolutely nothing special about the influence of lab measurement on behavior.

My view is expressed more compactly by Falk and Heckman (2009). They wrote:

> The issue of realism, however, is not a distinctive feature of lab versus field data. The real issue is determining the best way to isolate the causal effect of interest. (Falk and Heckman, 2009, p. 536)

They express concern that

> The casual reader may mistakenly interpret arguments about realism as an effective critique against the lab, potentially discouraging lab experimentation and slowing down the production of knowledge in economics and other social sciences. (Falk and Heckman, 2009, p. 536)

The idea that concerns about generalizability are not special to lab experiments is, in fact, found in Levitt and List (2007a) too, at the end of their paper:

> The sharp dichotomy sometimes drawn between lab experiments and data generated in natural settings is a false one. The same concerns arise in both settings regarding the interpretation of estimates and their generalizability outside of the immediate application, circumstances, and treated population.

Given the consensus among most experimental economists that realism, generalizability, or external validity are not especially important, it is useful to ask where that concern came from, historically. The term "external validity" was coined in the short landmark book on experimental design by Campbell and Stanley (1963) (cf. Bracht and Glass (1968)). The book was originally published as a chapter in the *Handbook of Research on Teaching*. Stanley was an educational psychologist. His interest in external validity probably originated in thinking about educational experiments—for example, what teaching methods produce the best learning under highly specific ecological conditions. (Stanley's view is the *policy view*.) Along these lines, Campbell and Stanley (1963, p. 5) note that

> While *internal validity* is the *sine qua non* [essential condition] and while the question of *external validity*, like the questions of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is

> obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of *known character* is the desideratum. (*italics* added)

An interest in "generalization to applied settings of known character" is therefore required, in order to care about and judge external validity. But this interest in applied settings of known character is the exception in experimental economics: Experimental tests "generalize to" the general behavior function assumed (by parallelism) to apply everywhere, not to any specific setting. But then there is no basis on which to judge or doubt its external validity. Of course, if the purpose of an experiment is to supply a policy-relevant answer to a particular external setting, then it is certainly important to ask about how well the experimental setting resembles the target external setting to judge its external validity.

These issues are brought into sharp relief by an example: Field data from high-stakes game shows with clear rules. Such data have been usefully exploited in many high-profile studies to learn about risky choice and strategic thinking (Andersen et al., 2008). But are game shows more "externally valid" than economic choices in canonical lower-stakes lab experiments? Not necessarily. As Bardsley et al. (2009) note:

> Such sources of data are often seen as "natural experiments." However, in this case the term "natural" should be interpreted with care. Although game shows are "natural" in the sense of arising independently of the research process, there are other respects, such as the nature of the decision problems and the presence of a presenter and studio audience, in which they seem quite untypical of normal life. Questions can therefore be posed about their broader external validity. (Bardsley et al., 2009, p. 26)

Game show field data *are* very useful, but not for their superior external validity compared to lab experiments. Game shows typically feature simple choices which can be analyzed clearly to judge whether participants are rational, for higher stakes than in the lab. So game shows are an excellent source of data on simple decision-making quality under high stakes . . . but game show decisions also have extreme audience effects, unusual (often opaque) participant selection, and imperfect control for participant comprehension.

Generalizability as Discussed by Levitt and List

Let's now turn to a discussion of generalizability as articulated in three papers by Levitt and List (LL) (2007a,b, 2008). One of their papers begins by saying that

> We can think of no question more fundamental to experimental economics than understanding whether, and under what circumstances, laboratory results generalize to naturally occurring environments. (LL, 2007a, p. 347)

LL "take as given that the issue of generalizability is important." Cited for support are the psychologists Pruitt and Kimmel (1977) and Rapoport (1970). Notice that all of those references are about 1970s experimental psychology, not about modern experimental economics. They also note a more subtle critique by Bardsley (2005).

Arguing *against* the importance of generalizability, they cite experimental economists Charles Plott, Vernon Smith, and Arthur Schram ("the 'classic' view of experimentation is one of theory testing, arguing that 'external validity is not an issue' in these instances" (LL, 2007a, p. 349)).

The "classic" view is sometimes called "blame the theory" by Bardsley et al. (2009). The idea is this: Since the theory as enunciated does not say that students in a lab behave differently from NYSE floor traders, for example, then an experiment which ignores that distinction is a valid test of a theory which also ignores the distinction. The phrase means that "the theory" is blamed for not including a role for student–trader differences. I do not like the phrase "blame the theory" because a person (or institution) must be blamed in order to make progress. The burden of hypothesizing (not "blame," per se) should rest somewhere. It would be useful for the economics profession to develop guidelines on where the hypothesizing burden should rest. Put concretely, if a critic says "Your experiment is not externally valid," this criticism should include content about what external validity is expected, why (is it written into the theory? should it be?), and what new experiments could be run to satisfy the critic's concern.

Comparing Methods for Empirical Inference

Following Falk and Heckman (FH) (2009), suppose that an outcome of interest Y is fully determined by a function $Y = f(X_1, X_2, \ldots, X_N)$ (an "all-causes" model, excluding pure error). Suppose we can establish causality in a lab or natural experiment by controlling one variable $X1$ and holding the other variables fixed or measuring them (denote them X^*). This vector X^* includes agent characteristics, endowments, incentives, rules, context, knowledge of other agents, and payoff structure. In lab experiment a vector X^* commonly—but not necessarily—consists of student subjects making abstract choices earning modest payoffs, who are carefully instructed about the relation between their choices and their payoffs. A common critique of the results of lab experiments is that the causal effect of $X1$ in the presence of lab X^* is different from behavior

> in a 'natural setting,' that is, for another set of conditions X^{**} including, for example, different institutional details, payoffs, and a different participant population. (FH, 2009, p. 536)

Falk and Heckman specifically note the apparent difference when X^{**} is sports card dealers (List, 2006) compared to lab experiments on reciprocity in offering higher product quality in response to higher prices. They note that

> If one is interested in the effect of social preference under a third condition X^{***} neither the undergraduate [lab] nor the sports-cards field study may identify the effect of interest. It is not obvious whether the lab X^{*} or the field X^{**} is more informative for the third condition unless a more tightly specified econometric model is postulated or a more precisely formulated policy problem is specified. (FH, 2009, p. 536)

That is, it is simply not clear that the lab conditions X^{*} or the card dealer conditions X^{**} are more like the "real world" if the particular target world is some X^{***}. This is the central issue: If the litmus test of "external validity" is accurate extrapolation to X^{***}, is the lab X^{*} necessarily less externally valid than the field setting X^{**}? How should this even be judged?

Since judging external validity is so slippery, let's return focus to control. There is little doubt that the quality of control is potentially very high in lab experiments (and measures of control quality—internal validity—are available). The usual analysis of field data typically has no direct control. Instead, econometric techniques and judgment are used to infer causality.

Ideal studies with field data exploit natural experiments in which there is either (a) true random assignment of agents to treatment groups (b) or a good instrumental variable that mimics natural random assignment (see Angrist and Krueger (2001)) to achieve control. However, the most convincing field studies with truly exogeneous instruments do not necessarily come from domains which are particularly "generalizable." Instead, wonderful instruments are often created by unusual quirks in policy or accidents of history. Their generalizability to other times and places is therefore in clear doubt (if skepticism about lab–field generalizability is applied uniformly). These doubts are accepted, as they should be, as a price to pay for unusually good natural–experimental control.

Note well that this logical claim is not at all meant to impugn the value of the most extraordinary quasi-experimental field studies. Instead, the point is just that the "best" field studies for establishing causality can be suspected of being difficult to generalize, just as the "best" method (experimental control) is suspected of the same weakness.[1]

Which methods most rapidly build up knowledge about $Y = f(X_1, X_2, \ldots, X_N)$? Each method has advantages and they have dynamic complementarities. Discoveries using one method are often then best explored further with another method. In the scientific view, the best empirical studies of all kinds (lab and field) seize on unusually diagnostic opportunities to measure $Y = f(X_1, X_2, \ldots, X_N)$ or to compare competing theories.

In comparing methods, lab experiments do have three special features that distinguish them: near-perfect replicability, measures of cognition, and obtrusiveness.

Because instructions, software, recruiting protocols, databases, and statistical tools are so easy to archive and reproduce, a lab experiment can be *replicated* very closely across a huge span of times and places. (And even the threat of replicability makes most experimental economists very careful.) In the lab it is also possible to

measure many aspects of *cognition* and neural behavior relatively easily. However, lab experiments are *obtrusive*: Because subjects must give informed consent (in the United States and many other countries), they know that they are in an experiment. Below we discuss further whether knowing you are being "experimented upon" has strong reliable effect on economic behavior that is different from comparable effects in parallel field settings. I think that it has little impact except in a few narrow settings.

How do field experiments compare to lab experiments on these important dimensions of replicability, measuring cognition, and obtrusiveness?

Replicability is typically more challenging for field experiments than for lab experiments[2] and is sometimes impossible. When access to a field site is controlled by a company, government, or NGO, new researchers could typically not replicate precisely what previous researchers have done.[3] Allcott and Mullainathan (2012) argue further that "partner selection" (of field site volunteers) is similar to subject selection in structure and perhaps impact on estimating effects. Furthermore, because field settings are naturally occurring, they invariably change over time (remote villages develop, firms go bankrupt, etc.) which undermines ideal replicability. The fact that field experiments are *not artificial*—and hence cannot be recreated over and over—then becomes an *impediment* to ideal replicability.

Field settings are also not typically ideal for measuring fine-grained cognition at the individual level (though it is often possible, in principle).

It is true that unobtrusiveness can be extremely good in naturalistic field experiments (such as List (2006)) if the subjects do not have to give informed consent (if their behavior is publicly observable). However, many field experiments are conducted in collaboration with a government or NGO (typically so, in development). In these settings, subjects may sense if there is a change in economic variables (such as a new subsidy) or that they are being experimented upon; they may also have a strong desire to please experimenters to get future opportunities.[4] If the worry about experimental obtrusiveness is that it produces special changes in behavior, it is possible that feeling "experimented upon" by a Western NGO in a poor country, for example, has a much stronger effect on behavior than being experimented on by a professor in a college lab. And in more developed countries, field experiments that are unobtrusive can inadvertently create attention to their treatments, as well as create backlash about unconsented participation, when some participants learn about their participation in a study (e.g., Milkman et al., 2012; Gelman, 2010).

Potential Threats to Generalizability

The notation $Y = f(X_1, X_2, \ldots, X_N)$ implies that whether a conclusion generalizes from one method and setting to another setting is the same as asking how different

the $X_1, X_2, \ldots, X_N$ variables are in the two settings and how sensitive the $f(.)$ function is to those differences. Conclusions are likely to generalize if all features of two settings are close or if $f(.)$ is insensitive to the features that differ. Note well that lab–field differences play no special role in this calculus of generalizability, except if experimental and domain features of X^* inexorably differ (the "alteration" that Bardsley et al. (2009) refer to).

It is true that in the past, many lab experiments in economics have tended to use a common configuration of design features. Typically, behavior is observed obtrusively, decisions are described abstractly, subjects are self-selected volunteers from convenience samples (e.g., college students), and per-hour financial incentives are modest. Let's refer to this type of experiment as the "common design."

It is crucial to note that the common design is not the only design. *All* of the common design features have been measured and varied tremendously in lab experiments.[5] If the goal is to extrapolate from conclusions to field settings that *do not* share most of these features, it is worthwhile to consider the threats to generalizability from the common design to such field settings. Doing so, Levitt and List (2007b) suggest the following:

> Behavior in the lab is influenced not just by monetary calculations, but also by at least five other factors: (1) the presence of moral and ethical considerations; (2) the nature and extent of scrutiny of one's actions by others; (3) the context in which the decision is embedded; (4) self-selection of the individuals making the decisions; and (5) the stakes of the game. (LL, 2007b, p. 154)

Note that the above sentence is equally persuasive if the word "lab" is replaced by "field." So the issue is whether factors 1–5 matter for lab behavior in a way that is never parallel to their influence in the field. Next we will examine the arguments and evidence about whether these features necessarily vary in the lab and field and how sensitive behavior is to these features in their special lab incarnation.

Presence of Moral or Ethical Considerations

There is no doubt that moral and ethical concerns trade-off with self-interest, in the lab and in the field. The issue for lab–field generalizability is whether lab settings provide systematically misguided evidence about the effect of moral and ethical concerns in comparable field settings. To be concrete, the issue is whether an experimental subject donating money in an artificial lab environment, knowing experimenters will see what they donated, necessarily creates moral and ethical forces which are fundamentally different from those at work, for example, when Sunday parishioners put money in a donation basket that is being passed around while other parishioners watch and donate, knowing that priests will add up all the donations later.

There is actually little evidence that sociality experiments are misleading compared to field settings *when matched with similar features*, and there is much evidence that such carefully matched lab–field results are close (see section entitled "Examples," below).

There are also, to no one's surprise, large differences in lab and field behavior when the two settings have *dissimilar features*. For example, in experimental dictator game allocations, when subjects are asked how much of $10 unearned income they would give to a stranger recipient, most give nothing; but a few give $5 and some give $1–2. The overall rate of giving is about 15% (Camerer, 2003, Chapter 2). A benchmark for giving in the field is the rate of charitable contribution as a fraction of income, which is typically around 1% in the United States. So it is clear that experimental subjects in many dictator games give a larger fraction of *unearned* income to *strangers*, when prompted, than the fraction of *earned* income people typically give in field settings to known *charitable causes*. But keep in mind that the dictator game was *not initially designed* to predict everyday sharing from earned income. Levitt and Dubner (2009) quote John List as writing

> What is puzzling, he [List] wrote, is that neither I nor any of my family or friends (or their families and friends) have ever received an anonymous envelope stuffed with cash.

The difference in giving in lab dictator games and in everyday charities (or anonymous giving to John List's family and friends) is no surprise because early lab dictator games were not designed to predict the level of anonymous giving. A more likely guide to field giving rates comes from lab dictator games in which money is earned. Indeed, Cherry et al. (2002) found that when dictators earned their money, giving in a $10 game was 4.3% of income, a number that is much closer to field giving rates.[6]

It is true that early discussions called the dictator game a way to measure "altruism," as compared to generous ultimatum offers by selfish proposers which are meant to strategically avoid rejection (e.g., Eckel and Grossman, 1996; Camerer, 2003). LL (2007a, Table 1) repeated this standard view, describing the "social preference interpretation" of dictator game giving as "altruism; fairness preferences, such as inequity aversion." The idea that dictator giving results from impure altruism, either warm glow or a preference for a social image of adhering to sharing norms, arose in the mid-1990s. The social image account was noted early on by Camerer and Thaler (1995), who used the phrase "manners."[7] Evidence accumulated in the last few years is also consistent with aspects of warm glow and social image motive (i.e., appearing to have good "manners").[8]

The social image or norm interpretation implies that allocations *should* vary with details of procedures and choice sets that affect image and norms. Indeed, the extreme control in the lab suggests that it is an ideal setting in which to learn about influences on sharing. The nature of entitlements, deservingness, stakes, and

obtrusiveness (how much you are being watched, by whom) can all be controlled much more carefully than in most field settings.

The Nature *and* Extent of Scrutiny of One's Actions by Others

A lab experiment is "obtrusive" (to use a conventional experimental term, see Webb et al. (1999)).[9] That is, lab subjects know that their behavior is being recorded and will be studied scientifically. The question is whether this obtrusiveness changes behavior in a way that is different than in a field setting with comparable obtrusiveness.[10]

The danger here is that being obtrusively watched by a lab experimenter creates demand effects which are unique to the lab, and can never be controlled for, if they were identified (an "alteration" problem; see Bardsley (2005) and Bardsley et al. (2009)).

It is *crucial* to note that evidence of obtrusiveness affecting behavior in *field* domains with no lab–field match (e.g., Alpizar et al., 2008; Bandiera et al., 2005) implies nothing about whether there is a *special* effect of experimenter observation which jeopardizes lab–field generalizability (i.e., that lab obtrusiveness has a bigger and more uncontrollable effect than field obtrusiveness).

To an economist, a demand effect is the result of a perception by subjects of what hypothesis the experimenter prefers, along with an intrinsic incentive subjects have to cooperate by conforming to the perceived hypothesized behavior. In this view the subject is an eager helper who tries to figure out what the experimenter wants her to do. (See Zizzo (2010) for a must-read thorough analysis of demand effects.)

First note that there is no clear evidence of experiment-specific demand effects of this type in modern experimental economics; there is merely suspicion.[11]

Why might strong reliable demand effects be rare in economics experiments? For strong demand effects to exist and to jeopardize generalizability, subjects must (a) have a view of what hypothesis the experimenter favors (or "demands") and (b) be willing to sacrifice money to help prove the experimenter's hypothesis (as they perceive it).

In most economics experiments involving choices, games, or markets, condition (a) is not likely to hold because subjects have no consistent view about what the experimenter expects or favors. Instructions are also very carefully written to avoid instilling any such expectations.[12]

Furthermore, in many economics experiments there are two or more competing hypotheses, and experimenters themselves are not sure which hypothesis will be true. (This is particularly the norm in behavioral economics experiments in which a rational hypothesis is often pitted against a behavioral alternative.) In these cases, even if the subjects did have accurate expectations of the range of what the

experimenters expect, it would not tell them what to do. Falk and Heckman (2009) suggest the following:

> It [obtrusiveness] is a minor problem in many experiments, especially if the decision environment is interactive and rich, such as in sequential bargaining or market experiments. Moreover, being observed is not an exclusive feature of the laboratory: many decisions outside the lab are observed. (p. 537)

Some evidence about condition (a), subjects' expectations of experimental demand, comes from Lambdin and Shaffer (2009). They replicated three classic experiments showing different types of preference reversal. Subjects chose between two outcomes in each of two treatment conditions. In each treatment condition the outcome consequences are the same, but the descriptions or choice procedures differ. (One example is the "Asian disease" problem showing the gain–loss framing effect.) Because the choices have identical consequences in the two conditions, it is possible that subjects might easily guess that the hypotheses have to do with ways in which choices would actually differ or not differ.

One group of subjects was shown both of the treatment conditions and asked whether they could identify the experimental research hypothesis. Subjects were quite confident that they could guess (mean 3.35 on a 7-point scale, with 1 as "extremely confident"). Eighty percent of the subjects said that the experimental hypotheses were "completely" or "somewhat" transparent. The authors also asked members of the SJDM society (who study questions like these); most of them (71%) thought that the hypotheses were transparent too.

However, the subjects' guesses about research hypotheses were only accurate 7%, 32%, and 3% of the time (across the three examples).[13] Thus, even very simple experiments that are the most likely to create demand effects spuriously producing experimental results do not necessarily do so, violating condition (a) for demand effects to matter.

For the sake of argument, assume that condition (a) does hold, and subjects accurately know that experimenters expect and favor a particular outcome. The next necessary step for demand effects to matter is condition (b), that the subjects will sacrifice earnings by behaving in a way the experimenter favors. Even if there is a desire to satisfy the experimenter's perceived demand (condition (b)), if that desire is simply a component of overall preference, then clear financial incentives should be able to overwhelm any such preference. A natural test for existence of a hypothesized demand effect is therefore to see if apparent demand effects shrink when more money is at stake. The fact that increasing stakes from modest levels to much higher levels typically has little effect (e.g., Camerer and Hogarth, 1999) suggests that as long as there is some salient incentive, demand effects are not playing a large role.

As a matter of logic and empirics, the case for clear, substantial demand effects in economics experiments is weak: Subjects are not necessarily likely to guess what

economics experimenters favor (because the experimenter favors no particular result, or because what they favor is opaque); and even if subjects do guess what experimenters favor, they may not sacrifice much in earnings to help the experimenter; and if they do sacrifice earnings to help the experimenter, we can remove those effects (and estimate their elasticity) by paying more. Thus, in principle there are methods to detect and alleviate demand effects with experimental design.

So where does the belief that strong demand effects exist in modern economics experiments come from? My hunch is that it comes from confusion between the canonical older methods of experimental psychology and modern methods of experimental economics. Those two systems of methods developed largely independently and are fundamentally different.

Some support for the confusion hypothesis comes from the historical sourcing in LL's writing. They assert that:

> Decades of research within psychology highlights the importance of the role obligations of being an experimental subject, the power of the experimenter herself, and the significance of the experimental situation. (LL, 2007b, p. 158)

To illustrate the dangers of demand effects in modern experimental economics, LL quote from an obscure short rejoinder made over 100 years ago by A. H. Pierce (1908) about how cooperative experimental subjects are. Pierce's comments were part of a debate about whether hysterical patients tell psychiatrists what they think the psychiatrists want to hear. LL also cite a classic 52-year-old paper by Orne (1962). Orne's discussion of demand effects was also partly motivated by his particular fear that hypnotized patients were overly cooperative (i.e., they respond to post-hypnotic suggestions in order to "help out"). The psychological and psychiatric sources that LL cite are interesting for distant historical perspective, but possible demand effects in studies of hypnosis and hysteria 50 and 100 years ago have nothing to do with modern economics.

LL also write:

> Schultz (1969, p. 221) described the lab as having a superior–subordinate relationship matched only by that of parent and child, physician and patient, or drill sergeant and trainee. (LL 2007b, p. 159)

Readers who have been either subjects or experimenters should judge for themselves how farfetched Schultz's comparisons are. If experimenters could guide the subjects' behavior so strongly and chose to do so, it would be a lot easier to publish papers!

To illustrate their fears concretely, Levitt and List (2007a) describe the subject's-eye view of (a fictional) "Jane," who participates in an abstract experimental game designed to test theories of firm behavior. Their discussion of obtrusiveness, subject naiveté, and demand effects in this tale simply rings false to an experienced lab experimenter. It bears little resemblance to typical subject

participation in the active working labs at places like Caltech and many other institutions with long-lived experimental economics research programs.

At one point, fictional Jane signs a consent form indicating that she

> understands her decisions are being monitored, recorded, and subsequently scrutinized for scientific purposes. Jane realizes that she is entering a relationship that has no useful parallels in her everyday life. (LL, 2007a, p. 348)

I disagree: The relationship that Jane is entering actually has *many* useful parallels in her everyday life.

First, if Jane has been in lab experiments before, then she *does* have a useful parallel—namely, her previous experimental experience. There are "no useful parallels" only if Jane is an experimental virgin.

Second, even if she is an experimental virgin, she will probably have been in many situations in which she makes abstract choices, given some simple rules, and her choices have economic consequences for her and will be recorded and analyzed by some people or system that created those choices and are interested in what she does. For example, Jane has done homework and taken exams. She took an SAT test in which her decisions are "recorded and subsequently scrutinized" for college admissions purposes. She probably played card games or board games with friends with abstract rules leading to wins or losses. She may have had job interviews, including written or oral exams, which are scrutinized for hiring purposes. She may have played a sport in which a coach scrutinized her performance, with the hope of being chosen for a team or given an award. All of these activities constitute potentially useful parallels in her everyday life for the experimental relationship.

The Context in Which the Decision Is Embedded

A major achievement of psychology, behavioral economics, and experimental economics (working in a loose consortium) has been to establish that contextual features and cues can have a substantial impact on behavior. Evidence of contextual effects has been building up for four decades, almost entirely from lab experiments in which context can be varied and all other differences can be controlled. LL go a little further and assert that "context matters and is not completely controlled by the experimenter" (LL, 2007b, p. 163). Context therefore makes their top five list of threats to generalizability of lab results.

Let's examine their argument closely: (1) LL argue that context matters (mostly) because experiments have conclusively shown that context matters by varying context and controlling other variables. Then (2) they fear that since context "is not completely controlled by the experimenter," lab contexts cannot be equalized effectively to comparable target field contexts.

The irony of their pessimistic conclusion, about lab context generalizing inconclusively to field context, is that premise 1 depends on strong experimental control

of contexts, while premise 2 depends on weak experimental control of context. Both cannot be right. If context with good control is achievable in the lab, then it should also be possible to create field-like context, if desired, in the lab too. Furthermore, if context is not *completely* controlled in an experiment, it is not any more controllable (or measurable) in typical field experiments or in naturally occurring data.

In any case, LL cite two types of data which are supportive of this hypothesis that uncontrolled context matters. Let's discuss one type of data, the artificial cross-cultural field experiments by Henrich et al. (2005) (in which I participated as a coach). They found that local sharing norms in analogous choices[14] (reported informally by the anthropologists) and cross-cultural ranked measures of cooperation and market integration themselves correlate with ultimatum offers. LL note that "the context the actors brought to the game and that experimenters cannot control—like past experiences and internalized social norms—proved centrally important in the outcome of play." It is true that these cultural sharing norms were not controlled by the experimenters, because the naturally occurring variation in norms was the independent variable of interest. However, variables associated with sharing norms could be *measured* to interpret different behavioral patterns (and correlated with ultimatum offers; see Henrich et al. (2006)). Just as in field experiments and analyses of field data, failing to completely control everything in a lab experiment is only a serious flaw if the uncontrolled independent variables cannot be measured and statistically controlled for.

Self-Selection of the Individuals Making the Decisions

Potential subjects are usually told a little about an experiment before they volunteer. Then there is self-selection that might make the resulting subject pool different from the recruited population and maybe different than in many field settings. (These are often called "volunteer effects.") One view about selection, based on old concepts in experimental psychology is that subjects are "scientific do-gooders" who "readily cooperate with the experimenter and seek social approval" (LL, 2007b, p. 165).

"In contrast," LL write, "market participants are likely to be a selected sample of individuals whose traits allow them to excel in the marketplace."[15] They suggest that self-selected experiment volunteers might be more pro-social than a random sample and less pro-social than "market participants." But Hoffman and Morgan (2011) found the opposite: There is more pro-sociality among workers in two internet industries (domain name trading and adult entertainment) compared to students. Several other studies have found that, if anything, students are clearly less pro-social than working adults (e.g., Alatas et al. (2009); Belot et al. (2010); Anderson et al. (2013); Falk, Meier and Zehnder (2013); and see Fréchette (2015 Chapter 17, present volume) on general subject pool differences).

In fact, there is some evidence that volunteer subjects are similar to nonvolunteers on relevant traits or behaviors. LL report (LL, 2007b, p. 166) that sports card sellers who volunteered for List's (2006) study were not significantly different in

reciprocity than nonvolunteers (as measured by their later behavior in an unobtrusive field experiment). Eckel and Grossman (2000) found that volunteers were *less* pro-social. Cleave et al. (2012) found no volunteer biases in a trust game and lottery choice task among $N = 1173$ students. Anderson et al. (2013) found no difference in pro-sociality among self-selected adults and a comparable sample of adult truck driver trainees who participate at a high rate (91%).

A likely reason why there are no systematic volunteering effects is that subjects volunteer to earn money, not because they are scientific do-gooders (as LL conjecture). In schools with active economics labs, subjects *do* see themselves as market participants whose traits allow them to excel in the marketplace for experimental labor. Students work in experiments as temporary employment, earning money for interesting work with flexible hours.

Finally, since self-selection into market activity *is* an important force in economic behavior, a question arises about how selection can be studied most conclusively. Ironically, while unmeasured selection into experiments is a potential threat to generalizability, controlled lab experiments provide some of the best possible methods for learning about self-selection. The ideal approach to measuring self-selection is to measure characteristics of a large pool of subjects, allow them free choice into different experimental conditions, and then see which subjects choose which condition and how they behave (see, e.g., Lazear et al. 2012). For example, suppose you wanted to test whether market participants are a self-selected sample of individuals whose traits allow them to excel in the marketplace. An ideal way to do so is to measure their traits, allow them to select into different marketplaces, and see who selects into what markets and whether they excel. This is typically difficult in field data and extremely easy in the lab.

Indeed, a large study by Slonim et al. (2013) gives important new evidence on selection characteristics into experiments, in general. They asked subjects in economics tutorial sections to voluntarily answer several demographic, behavioral, and preference questions. Almost everyone (97%) complied, yielding a large sample ($N = 881$). Later, those students could volunteer for economics experiments. Slonim et al. can therefore compare characteristics of those who self-selected into, or out of, the experiments.

The self-selected participants spent less money and worked fewer hours each week, were more likely to have a high IQ (measured by CRT) with more consistent preferences, were disproportionately economics/business majors, and saved more money. These participation effects are all consistent with a profile of volunteers who are looking for flexible part-time work which is intellectually interesting, perhaps about economics. (This profile is a far cry from LL's analogies of experimental subjects to medical patients, military trainees, and scientific do-gooders.)

Another small participation difference is pro-sociality. One out of five measures of pro-sociality was higher for experimental participants: They volunteered time to charities more frequently than did nonparticipants ($p < .05$), but hours of volunteering, frequency, and amount of money donation were not different.[16]

Participation biases like these certainly must be considered in interpreting experimental results. An advantage of unobtrusive natural field experiments is that we choose the subjects[17]; in typical lab experiments, the subjects choose us. However, there are at least three possible ways to partly correct for participation bias. The first is to estimate treatment effects for a particular characteristic within the selected population (e.g., whether self-selected women behave differently than men in the experiment) and then extrapolate that result to the special characteristics of the entire population of interest in making a guess about population general behavior. This shortcut assumes that self-selected women behave like nonselected women, which could be wrong. Two better methods simply increase the participation rate. The subjects-as-workers model implies that it is usually possible to raise participation close to 100% by paying large enough participation fees. This is simply an expensive way of reducing participation bias. A third, related method is to study populations which are motivated or compelled to participate at high rates for other reasons, such as in businesses or military organizations, or add experimental questions to representative random samples that are "rentable" from survey organizations (and whose participation decision is already well understood). Finally, note that challenges from selection bias occur in all fields. They are not special to experimental economics and are not especially insurmountable. Survey researchers and medical researchers, for example, have developed good methods for correcting such biases. Therefore, while self-selection will typically be present in economics lab experiments, it can be present in field experiments too, and correcting for mistakes in generalizable inference which result from self-selection is not impossible and will sometimes be easy.

The Stakes of the Game

The possible influence of economic incentives ("stakes") on behavior has been explored, many, many times in experimental economics. Smith and Walker (1993) examined about two dozen studies and concluded that there is support for a model in which higher stakes motivate more cognitive effort, reducing response variance and bringing responses closer to rational ones. Camerer and Hogarth (1999) look at a larger sample of 74 studies. They conclude that the largest effects of incentives come from comparing hypothetical payoff studies (with zero marginal financial incentive) to those with some incentives. The most reliable effect is that noisy response variance falls. In some cases (e.g., judgment tasks where there is a normatively correct answer) higher incentive induces more rational choices. Incentives also seem to reduce effects of socially desirable behaviors (e.g., there is less dictator giving, and less risk taking, with higher incentives).

Most studies in foreign countries where purchasing power is low have found that the basic patterns in a variety of games and markets which are observed at modest stakes in traditional subject pools also replicate when stakes are very high. Therefore, the effect of stakes is no longer a sound basis for critiquing lab–field generalizability.

It is also frustrating that the common assertion that raising stakes will change behavior is almost never expressed in the form of how *much* behavior will change in response to a change in stakes. It is true that in some field settings with very high stakes, it may be impossible as a practical matter to conduct lab experiments with matched high stakes. But there is also little evidence that substantial differences in stakes (within ranges that have been varied experimentally) change conclusions derived from modest stakes.[18]

New Experiments Are the Best Remedy for Concern About Generalizability of Older Experiments

Falk and Heckman (2009, p. 537) note that

> Ironically, most objections [concerning lab evidence] raise questions that can be very well analyzed with lab experiments, suggesting the wisdom of conducting more lab experiments, not fewer.

To illustrate Falk and Heckman's point, recall LL's list of candidate factors that limit lab generalizability: Moral considerations, scrutiny, context, self-selection, and stakes (LL, 2007b, p. 154). The most conclusive evidence that variation in these factors changes in behavior *comes from the lab itself*. The lab evidence is the most conclusive because field measures (and natural experiments) have much weaker control over those factors than lab environments do.

Above I discussed how lab experimental control might prove especially useful in studying the effect of self-selection (although it has not been used very often so far). Another example is scrutiny (obtrusiveness). The important part of obtrusiveness that we care about is how it is perceived by the economic agent. Measuring obtrusiveness requires us therefore to vary its level and nature *and* to accurately record how it is perceived by the agent. In naturally occurring field examples, obtrusiveness might well vary (e.g., supervision of worker effort) but the variation would typically be highly endogeneous from firm choice—hardly exogeneous—and often very difficult to measure. There are beautiful examples of such variation in field experiments (e.g., Rebitzer, 1988), but a high degree of control is hardly ever attained in typical field data without quasi-experimental control.

Examples

So what do the best available data about lab-field generalizability say?

I will focus mostly on two kinds of data. The closest lab–field comparisons are carefully designed to have the features of particular field settings and compare the behavior of the same types of subjects in the field and in the lab facsimile. Unfortunately, there are a small number of these close comparisons (only six). There are many more examples of lab experiments which measure a behavior (such as prosocial cooperation) in one subject pool, as well as a related field behavior which is

not closely matched in structure, but which is expected to be correlated with the lab results at an individual or group level.

The Closest Tests of Lab–Field Generalizability

Sports Card Trading

The best series of closely comparable lab and field experiments uses transactions with sports card dealers, conducted over different years and compiled in List (2006).

The setting is sports-paraphernalia dealer conventions. Dealers arrive and spread their goods for sale on tables. Buyers mill around, often request a particular quality level (for cards which can be independently graded) and offer a price. The crucial question is whether dealers offer items of the requested quality or give lower-quality items.

For brevity, let's focus on the lab experiment "lab–market" which most closely matches the economic environment in three of the field experiments. Five other treatments are interesting but are not such close matches. List (2006) discusses these in detail. For the purpose of this essay, comparing field experiment conditions that are much different from their lab equivalents is just of much less interest than comparing two lab and field conditions that are matched closely by design.

The transactions are simple: Buyer subjects are recruited (for a flat participation fee) to approach actual dealers and offer either $20 requesting a card with quality 4, or $65 requesting a card with quality 5. Note that these experiments are unusual, by experimental economics, because the buyers have no financial stake in the outcome.

One key feature of the study is that the cards which are actually bought by subjects can then be independently graded to compare the "actual" (graded) quality with what was requested. The experiment takes place near the field marketplace and is closely matched in protocol. There are two key variables in the market of interest (given that prices are fixed by experimental control): First, how much quality is supplied by dealers? And second, does supplied quality respond to price?

The first finding is that there is a clear drop in quality from the lab to the field. The average quality levels for the $20 and $65 cards are 3.1 and 4.1 in the lab and are 2.3 and 3.1 in the field. So the shortfalls in quality from what were requested are around one quality point lower in the field. All dealers seem comfortable delivering less quality in general in their naturally occurring environment (when they do not know if quality is later measured, which they do know in the lab setting).

The second finding is that the response of quality to price is closely comparable in the lab and field. Table 14.1 reprints most of the statistics that appear in List's Table 4. The lab–context experiment has a price coefficient of 0.05, but its larger magnitude could be due to the fact that buyers are incentivized for performance in that treatment (and are not incentivized in the others reported).

Table 14.1. Coefficient from Tobit Regression of Quality on Price and Estimated Gift Exchange Coefficient (List, 2006)

	Lab-Context	Lab-Market	Field	Field (Pre-grading)	Field (Announce Grading)	Field (Grading)	Field (Pooled)
	Described as Cards	Cards	Cards	Tickets	Tickets	Tickets	Tickets
Within-subjects data?		Yes	Yes				
Price coefficient	0.05	0.02	0.02	–.001	0.02	0.02	0.02
	(4.3)	(4.4)	(6.6)	(0.01)	(2.1)	(1.1)	(2.6)
Gift exchange estimate θ	$0.65	$0.45	$0.21	$0.01	$0.17	$0.23	$0.19
	(4.7)	(2.1)	(5.0)	(0.3)	(1.1)	(1.1)	(2.3)
Buyers incentivized?	Yes	No	No	No	No	No	No
N	32	60	100	60	54	36	90

Note: *T*-statistics in parentheses. Dealer random effects included (except for Lab-Context). The column headings above correspond to the original (List, 2006, Table 4) headings of: Lab-Context, Lab-Market, Floor (Cards), Floor-Announce Grading, Floor-Grading. Original Columns Lab-R, Lab-RF, Lab-RF1 with Modest N (25–27 each) are omitted; readers are encouraged to see List (2006) for details.

In the other treatments with no buyer incentive the coefficient of quality on price is always 0.02, except for ungradeable tickets (the coefficient is –.0001). This implies that paying an extra $50 will increase provided quality by about one unit. The statistical strength of this gift exchange effect is actually *stronger* in the field card market than in the lab markets (due partly to larger sampling in the field).

Table 14.1 shows good lab–field generalizability. Effects of price on quality in the lab and the field (the "price coefficient") are exactly the same (except for no grading). The financial return from gift exchange (θ) is about twice as high in the lab–market condition, but there is no significant difference between that value (= 0.45) and those in other treatments, using an appropriate difference-in-differences test.

In *Science* Levitt and List (2008) wrote:

> Some evidence thus far suggests that behavioral anomalies are less pronounced than was previously observed in the lab (List, 2006, Figure 1). For example, sports card dealers in a laboratory setting are driven strongly by positive reciprocity, i.e., the seller provides a higher quality of good than is necessary, especially when the buyer offers to pay a generous price. Yet, this same set of sports card traders in a natural field experiment behaves far more selfishly. They provide far lower quality on average when faced with the same buyer offers and *increase quality little* in response to a generous offer from the buyer. (List, 2006, p. 910) (italics added)

Hmm. This paragraph—and especially the italicized phrase—does not describe the full results, as summarized in Table 14.1. What is going on? The discrepancy is because the *Science* reference to "sports card traders in a laboratory setting" refers to the smaller subsample (30%), nonlocal dealers only, not to the full sample.There is a substantial difference in behavior between "local dealers" (who usually go to the conventions repeatedly and may operate a storefront business or website) and "nonlocal dealers," as self-identified by surveys. Their conclusion refers only to the nonlocal dealers and ignores the larger group of local dealers.

Figures 14.1 shows the average quality for both nonlocal and local card dealers in the lab and field experiments. (Figure 14.1a is the one reported in LL (2008) with standard error bars approximated). There is apparently a sharp difference between responsiveness of local and nonlocal dealers in the field, but no such difference in the lab.

The local–nonlocal comparison is indeed important, because the reciprocity exhibited by the locals *could* be driven by reputational concerns (for local dealers only), by social preferences, or by both. If the nonlocals have lower reputational concerns, their behavior does provide better isolation of pure social preference effects.

A fresh look at raw the data[19] shows two new observations not reported in the original paper.

Any lab–field gift exchange difference will be revealed most powerfully using a within-subjects comparison (unless the within-subject sample is too small). In fact, there are a small number of dealers who participated in *both* the lab–market and field card treatments. This comparison deserves special attention because: (a) it is the closest match of lab and field features and (b) the within-subject comparison has more power to detect lab–field differences. Other lab–field comparisons do show larger differences in reciprocity but also have more uncontrolled design differences than this single lab–field comparison for which within-subject comparison is available.

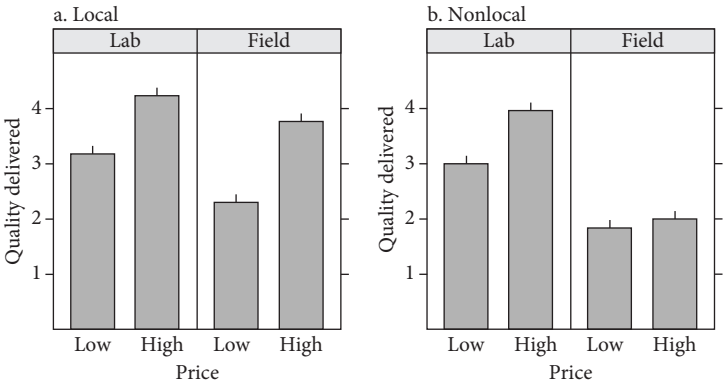


Figure 14.1. Quality responses to low and high price offers. (a) Local dealers (lab, left; field, right). (b) Nonlocal dealers (lab, left; field, right). Adapted from List (2006).

Table 14.2. Mean Quality Supplied in Response to Higher Prices (Reciprocity) in Adjacent Lab and Field Markets by Dealer Sample

	Non-local		Local	
	Mean (std. dev.)	*N*	Mean (std. dev.)	*N*
Lab	1.11 (1.17)	9	0.95 (1.32)	21
Field	0.13 (0.92)	15	1.40 (0.81)	35
Nonparametric Tests: statistic, two-Tailed *p*-Value				
Mann–Whitney	38, 0.084		461.5, 0.114	
Fligner–Policello	2.09, 0.020		1.45, 0.073	
Epps–Singleton	3.83, 0.429		22.07, 0.0002	

The key data for each dealer are the differences between the quality supplied for the $20 card and the $65 card, in the lab and field treatments. The important comparison is whether that quality difference is larger—more reciprocal—in the lab than in the field.

These results are summarized in Table 14.2. The nonlocals offer 1.11 more quality units in the lab for the more expensive card. The nonlocals only offer 0.13 more in the field, so they are indeed less reciprocal in the field than in the lab. However, the locals are somewhat *more* reciprocal in the field (1.40 compared to 0.95). **That difference is the first new observation.**

In List (2006) and Levitt and List (2008), the difference in reciprocity (i.e., the seller's quality response to the buyer's offered price) between lab and field was not tested statistically. The proper test is a difference-in-difference test, in which the quality differences for high and low prices are compared in the different lab and field locations. Because the distributions of quality differences are discrete and not at all Gaussian, nonparametric tests are appropriate. Three different tests give different results (Table 14.2. One test (Fligner–Policello) shows a strongly significant negative difference for the nonlocal dealers (less field reciprocity) and a weaker positive effect for locals. Another test shows the opposite conclusion about significance, more strongly (Epps–Singleton). Overall, the conclusion that the lab and field show different reciprocity (especially for nonlocal dealers) is suggestive but is just not robustly significant (in two or more of the three tests). **That's the second new observation.** To be crystal clear, I am not sure whether dealers behave differently in the lab and field, in the most closely matched comparison of lab–market and field (cards). This is not the same as concluding that their behavior is the same.

While actual purchases are the revealed-preference data of most interest, List (2006) also records cases in which a seller claims what quality they will supply and then compares that claim to what is actually offered. Overclaims that are higher than what is delivered are called "mendacious." Dealers only make claims in the lab for a minority of trades (16%), so the within-subject comparison between lab and

field mendacity has too little power to draw a conclusion from that comparison. However, mendacious overclaims are clearly more common in the field.

The conclusion about quality delivery in the lab and field from this within-subject comparison is certainly more equivocal than Levitt and List (2008) claim. Their claim is numerically correct: The average quality difference of 1.11 in the lab is greater than 0.13 in the field. But the test is underpowered and is not statistically strong. However, List (2006, p. 32) concludes:

> The data suggest that social preferences do not have a major impact in these particular markets.

This conclusion goes in the wrong direction for the local dealers (70% of the sample), and it is not statistically robust for the nonlocal dealers (the other 30%). Keep in mind that this study *was* clearly well-designed to identify a possible failure of lab–field generalizability, but there is no statistically robust failure. Furthermore, while the dealers do generally offer lower quality in the field, the fact that the dealers offer *any* good-quality cards in response to price offers,[20] as compared to offering the lowest quality for all offers, is evidence of some degree of reciprocity or other social preference.

Finally, this study is an important one because it has led to some broad, widely publicized conclusions about human nature. A news article (Stanley, 2011) concluded that List's (2006) card-dealer paper "helped refute a prevailing theory—which behavioral economists had arrived at in the lab—that humans were altruistic by nature." The article then quoted List as saying "Very few people will not screw each other. There are very few nice people out there."

(By the way, these conclusions about human altruism conflict sharply with the prevailing view of human exceptionalism in biology and anthropology. Scientists in those fields have amassed much evidence that humans are more pro-social toward non-kin than all other species. For example, Boyd and Richerson wrote: "Humans cooperate on a larger scale than most other mammals. Among social mammals, cooperation is mainly limited to relatives. . . . The scale of human cooperation is an evolutionary puzzle" (Boyd and Richerson, 2009, p. 3281).[21]

Open Air "Flea" Markets

List (2009) conducted a complex set of experiments about "open air" flea markets, which is something of a sequel to List (2006). In these markets a vendor rents a venue and charges sellers a fee to set up a booth to sell both standardized and unstandardized goods, for a day or weekend. Buyers show up, and they browse and haggle. List builds a staircase of institutional steps from the field setting, adding tighter experimental control at each step. The project exploits information from an unnamed "mole" informant seller who is colluding with other sellers to fix prices by an agreed-upon markup from the marginal cost paid to a common middleman.

In the lab and framed field experiments, buyer and seller values and costs are induced in the classic Smith–Plott design. In a natural field experiment, confederate buyer values are induced and the seller marginal costs are assumed (from the mole's inside information). Because the mole described clear price collusion, the data analysis is largely focused on whether there is cheating by undercutting the collusive price (as well as other related statistics such as price and allocative efficiency).

The closest lab–field designs are a framed table treatment in which worthless prop goods are traded for induced values, as compared to a natural field treatment in which actual items are traded (for the induced buyer value and for assumed seller marginal cost).

The key sample is 17 sellers who participated in either lab or field treatments, and also in the natural field treatments. While lab and field behavior are not always closely associated, List notes (without giving detail) that

> the best predictor of whether they will cheat in the natural field experiment is their measured cheating rate in the framed field treatments and the lab treatment with context (List, 2009, p. 44)

The overall deviation of transacted prices from collusive prices is 12.8% in framed table and 19% in natural field. Allocative efficiencies are 77% and 90%, respectively. Not enough details are reported to judge whether these modest differences are significant or not. One can therefore conclude that there is no reported evidence of significant differences in the lab and field in this chapter, though there are small behavioral differences.

Donations to Student Funds

Benz and Meier (2008) compare naturally occurring donations of CHF 7 and CHF 5 to two student funds (gleaned from student records), with experimental donations after a class from a CHF 12 endowment to the same two funds.

First note that even with this careful control, there are substantial differences in the lab and field treatments. The lab experiment was done at the end of a class, subjects were endowed with money to donate, and they could donate amounts in increments of 0.5 CHF. The field donations were done from their own money and were all or nothing.

The average donation in the lab experiment was 9.46 CHF, compared to field donations of 9.00 and 9.50 in the four semesters before and after the lab experiment. The bad news is that the within-subject correlations between lab and field donation are modest, from 0.22 to 0.31.

In any case, while the lab–field correlation at the individual level is modest, the average donations in the two cases are very close. (In their 2007b paper, LL noted the low lab–field correlation across subjects, but did not mention the very close match of overall donations in the two conditions.) Is 0.30 a small or big lab-field correlation? Between-situation correlations of trait-like behavior are often

not much larger than 0.30. There may be a practical upper limit on how much lab-field consistency we expect within people, *but the same limit applies to field–field comparisons*. Unique evidence on this comparison is provided by Liebbrandt (2011), who collected several lab measures and field measures of prosociality among Brazilian fishermen. He found that different lab and field measures of pro-sociality correlate modestly (around $r = 0.30$), and field–field correlations are similar in magnitude to lab–lab and lab–field correlations.

Soccer

Palacios-Huerta and Volij (2008) (PHV 2008 herein) designed a 2×2 lab matching pennies game designed closely to resemble soccer penalty kicks, in which players appear to approximately randomize (Palacios-Huerta, 2003). In their simplified lab game a kicker and goalie choose left or right simultaneously. The game payoffs are kicker win rates estimated from 10 years of actual data from pro games.

PHV found that professional players approximate the mixed strategy equilibrium prediction in the lab games, with some modest deviations and serial correlation.[22] College student subjects with no soccer experience deviated from equilibrium mixtures and had too many runs (a common pattern). However, their lab data provide no evidence at all of lab–field generalizability because they have no relevant "field experience." Interestingly, a sample of students who were playing in a serious amateur league played the lab games much like the professionals, which suggests that field experience could create generalizable skill in competitive randomization (cf. Martin et al. (2014) with chimpanzees). The data from professionals in this experiment—which is the only group for whom lab and field data are both available—show a close match of play in the field and the lab (though cf. Wooders (2010)). (Student lab behavior and professional field behavior are indeed different, but this comparison confounds lab–field and experience; when the confound is broken by focusing only on the professional lab–field comparison, there is a reasonable match.)

Levitt et al. (2010) report evidence that poker, bridge, and American-league soccer players do not randomize as mixed equilibrium predicts. But they have no field data from those subjects, so their results provide no direct evidence lab–field generalizability.

Communal Fishing Ponds

Stoop et al. (2010) did a meticulous set of experiments on fishing and public goods contribution. Their baseline field setting is ponds in Holland where fishermen each pay a private fee and then have a communal pond stocked with a specific number of extra fish that anyone can catch. They conduct a field experiment in which the incentives are changed. In their baseline condition the pond is stocked with 38 fish for 16 subjects, and they can only catch 2 fish per fisherman (in a 40-minute period). In a voluntary contribution (VCM) condition, the other fishermen in a 4-person

group share a 6 euro bonus for each fish a specific subject catches under their 2-fish quota. They record both catches and fishing effort (measured by the number of casts of their fishing lines per minute). The VCM condition has no effect on behavior in the field.

The remarkable part of the paper is that lab experiments are created with similar financial incentives (and contextual language about fishing). Fishermen who participate in the field setting, as well as students, are subjects (in separate groups) in the lab experiment. They find that fishermen are actually more pro-social in the lab, and they are even more pro-social when the lab structure is conducted outside—that is, at the fishing pond, rather than in the lab.

There is clearly a difference in fishermens' pro-sociality in the field setting (where VCM incentive to underfish do not change behavior) and in the lab (where financial incentives do lead to more cooperation, even moreso than students). While this is the best evidence that pro-sociality is weaker in "natural" field settings than in lab analogues, it is not strong evidence because there are many differences between field and lab. In the field an experimental period is 40 minutes. Fisherman make about three casts in five minutes. So it is easy for others to see very rapidly whether their peers are cooperating by casting less, which might make cooperation dissolve quickly. In the lab a period is a couple of minutes and involves no real-time observation of what other people are doing. Underfishing in the field, to create more euros for peers, also means sitting idly by while others might fish. As result, there is a possible "boredom cost" which is not induced in the lab equivalent setting.

This study should be seen as an example of how lab and field pro-sociality could differ within a single subject pool (since the same Dutch fishermen participate in both types of experiments), but is also a cautionary tale about how many features of the two kinds of experiments need to be equalized to permit sharp lab–field comparison (there is a lot of slippage here in the lab–field match).

Proofreading and Exam Grading

Armentier and Boly (2013) did a beautiful simple study on responses of actual effort to incentives and monitoring in the lab and field. Subjects were instructed to find spelling errors in papers typed in French from dictated tapes. In the monitoring conditions, either one or five out of 20 papers were examined and subjects were penalized if they miscounted the errors, based on the absolute deviation of detected and actual errors. In a high-wage condition, they were paid more (but no performance-linked pay). The lab subjects were students at CIRANO in Montreal. The field subjects were recruited in Burkina Faso (mostly students at the local University of Ouagadougou) and told they were spell-checking "dicteés" (exams for civil service jobs). Note that the Ouagadougou subjects did not know they were in an experiment, but the Montreal lab subjects did.

Despite the large difference in subject background, the effects of monitoring and higher pay are generally identical up to two decimal places. The GLS

coefficients of low and high monitoring on absolute deviations are –1.31 and –1.38 in the lab and –1.33 and are –1.49 in the field. There are also very close "period" effects (the order in which the monitored papers were done) on a slowdown in effort. And females are better graders in the lab (coefficient –.71) and the field (–.75). In a related report, Armantier and Boly (2013) measure corruption rates by inserting a bribe into one of the exams. They find very close rates of corruption (acceptance of the bribe and grade misreporting) in the student and worker groups.

Fungibility of Cash and In-Kind Subsidies

Abeler and Marklein (2010) did parallel field and lab experiments on fungibility of money. They compared overall spending and beverage spending when diners at a German wine restaurant were given identical-value 8 euro vouchers for either a specific good (beverages) or an entire meal. Almost all diners spend at least 8 euros on beverages, so the restricted beverage voucher is essentially as valuable as the unrestricted voucher. They find that subjects spend more on the targeted goods in the field restaurant setting and in a stylized lab setting (with induced value consumption). The design does not permit a comparison of effect sizes, but is clear evidence that the sign and significance of the specific-voucher effect is similar in field and lab.

Money Sharing (Dictator Games)

Stoop (2014) gave Dutch subjects a postmarked partly transparent envelope with two 5-euro bills and a thank-you note to Tilburg University volunteers. Subjects could keep the money, or mail it to the volunteer (perhaps after taking out one 5-euro bill). The study carefully applied similar protocols in four treatments: to a student lab group, two representative-sample lab groups (participants in one of the groups take the envelope home before deciding whether to mail it)), and a field group who found the envelope in their mailbox. Stoop painstakingly controls for possible differences across treatments.The results show no lab–field differences across this "bridge." Subjects either keep all the money or send back all 10 euros. The percentages of those sending back the money range from 57% to 45% across treatments.

Summary

These studies on sports cards, markets, donations, fishing, grading, and restaurant spending provide the closest matches of lab and field settings, protocols, and subjects that have been conducted. Only one—fishing—shows poor generalizability of the essential behavior from the lab and field, which is statistically reliable. It therefore appears, empirically, that when lab and field dimensions are carefully matched, good generalizability is the rule and poor generalizability does occur, but is the exception. This pattern suggests a tentative *general* conclusion about lab generalizability:

Claim: There is no replicated evidence that experimental economics lab data fail to generalize to central empirical features of field data (when the lab features are deliberately closely matched to the field features).

This bold claim is meant to accomplish two goals: First, the default assumption in the economics profession should be that lab experiments *are* likely to generalize to closely matched field settings, until more clear studies show the opposite (and empirically show typical conditions for poor generalizability). Good generalizability is the default assumption and is generally supported by direct comparisons, in other fields such as biology, which routinely compares animal behavior in the lab and in the wild.

Second, the claim is phrased boldly to attract the attention of people who think it is wrong. The claim is an assertion about empirical facts. Readers are free to say they don't believe it, are unconvinced, or remain skeptical, but it is a claim that is currently either True or False (depending on judgments of what "deliberately closely matched" means). Skeptics should either bring existing examples—with close design matches—to the attention of the profession, or do more studies with close design matches to add more results.

A different claim is whether there will *ever* be any evidence of poor lab–field generalizability (using a close design match). The answer is certainly Yes, thanks to Stoop et al. (2010).

Apparent Good Generalizability with Imperfect Lab–Field Design Match

In a few other studies, details of the lab design *were* specifically chosen to be highly comparable to a field setting. However, they are less conclusive than the studies in the previous section because either the subject pools are different or there are no field data on lab participants. In these studies, the lab–field match in effect size and direction is generally good, even though the lab environments have some of the features hypothesized to undermine generalizability.

Sports Good and Consumer Good Trading

List (2003) measured the strength of endowment effects in simple exchange experiments with sports paraphernalia traders (including professional dealers). The field data are artificial experiments conducted with participants who are endowed (randomly) with either good A or B and asked if they would trade it for the other good. Experienced traders and dealers trade a little less than half the time (45%), and inexperienced traders rarely trade (10–20%). There is a strong tendency for market experience to reduce endowment effects.

However, there are substantial endowment effects in *n*th-price auctions of sports card goods in field settings among dealers and nondealers. There appears to be an endowment effect even among dealers because the mean selling and buying

prices are $8.15 and $6.27, a ratio of 1.30 (comparable to other ratios, e.g., Bateman et al. (1997); though the ratio for nondealers is much higher at 5.6). The WTA–WTP difference is not significant ($t = 0.87$) but such a test does not have much power to find a typical empirical effect (with $n = 30$ dealers in each group). Indeed, List (2003, footnote 11) says "I do not consider the point estimates herein to fully support neoclassical theory" (on the grounds that implied income elasticities for the endowed goods are "implausibly large").

The paper also briefly reports a four-week experiment in which students come to a lab once a week, and each time are endowed with one of two different goods then allowed to trade for the other goods (with different goods pairs each week). In the first week, only 12% trade away their endowed good, and 26% do in the fourth week. This trend is comparable to the experience effects that are associated with more trading in the field experiments.[23]

The strong lab effect implies that the title of the paper, "Does Market Experience Eliminate Market Anomalies?", is incomplete because it does not refer to the lab data at all. A more informative title would be, "Does *Lab or* Market Experience Eliminate Anomalies?" The difference in language is important because the paper is typically cited as showing that field evidence from experienced market traders overturns an effect that is well-documented in simple lab settings. This rehearsed encoding of the result can easily lead to the mistaken impression that there are endowment effects in the lab but not among experienced traders in the field. That is certainly not the full conclusion since experience reduces endowment effects in *both* lab and field.

Semantics and titling can matter a lot since people remember the title or the gist of the paper better than its detail (if they even read it all), since gist retention is how unaided memory typically works. In fact, none of the more widely cited papers[24] which are reported as citing List (2003) mention the student lab data *at all*. For example, Haigh (2005, p. 524) write:

> In light of some recent studies (e.g., List, 2002, 2003, 2004) that report market anomalies in the realm of riskless decision-making are attenuated among real economic players who have intense market experience, the current lot of experimental studies and their support of MLA [myopic loss aversion] may be viewed with caution.

Note that their passage distinguishes between "real economic players" and "experimental studies," as if the results are fundamentally different in those groups. But the List (2003) paper actually shows behavior that is fundamentally *the same* in those two groups.

Event Partitions in Prediction Markets

Sonneman et al. (2011) studied whether a simple bias in probability judgment is manifested in lab experiments, field experiments, and field data. The bias is that when a continuous variable is partitioned into N sets, judged probabilities over sets

tend to be biased toward 1/*N*. For example, if the outcomes of the NBA playoff finals, measured by total games won, are partitioned into (0, 3), (4–6), (7+), the combined judged probability of the first two sets is larger than the judged probability of the "packed" set (0–6) when the partition is (0–6), (7+). A common intuition is that self-selection into field markets, stakes, and experience could reduce or erase these types of psychological effects.

To test this hypothesis, Sonneman et al. (2011) compared two-hour lab markets for naturally occurring events, seven-week field experiments on NBA and World Cup outcomes, field data from 150 prediction markets for economics statistics, and data on betting odds for a million horse races (Snowberg and Wolfers, 2010).

Some partition dependence is evident in all four markets. Statistical analysis indicates that the probabilities implied by economic statistics market represent a mixture of a 1/*N* prior belief with judged likelihood, much as in lab experiments. The combination of trader self-selection into those markets, its special context, and incentive does not undermine the generalizability of the lab result.

Sharing with Charity

An experiment on dictator game sharing with a specific Chinese poverty charity showed a similar drop when the income endowment was earned versus unearned in a student lab experiment versus citizen customers outside a supermarket (Carlsson et al., 2013). The students did give more in general, but that effect is confounded with lab–field, demographic differences, and some procedural variables. This is an illustration of Kessler and Vesterlund's (Chapter 18, this volume) important point that the lab and field could generate different behavioral level effects, but also generate comparable comparative static responses to design variables.

Swedish Lotteries

Östling et al. (2011) collected data from a daily Swedish lottery in which players choose integers from 1 to 99,999, and the lowest unique integer wins a large prize. This *l*owest *u*nique *p*ositive *i*nteger (LUPI) game provides a rare sharp test of mixed strategy equilibrium in the field. The Poisson–Nash equilibrium predicts comparable numbers of choices of numbers 1–5000 with a sharp dropoff after that point. Compared to that benchmark, there are too many low numbers, not enough numbers in the range 3000–5000, and too many higher numbers. The data are fit reasonably well by a QR cognitive hierarchy model with average thinking $\tau = 1.80$ (close to a typical estimate from many lab experiments; e.g., Camerer et al. (2004)).

Östling et al. (2011) then designed a lab experiment to replicate the key theoretical features as closely as possible while scaling down the number of players and the integer range. The central empirical feature of the field data—too many low and high numbers—also occurs in the lab data. This is an example of when qualitative results are a close match but quantitative details are not.

Silent Auctions

Isaac and Schnier (2006) compare field data from three "silent auctions" with lab experiments designed to be closely comparable. They focus on the frequency of "jump bids" which top previous bids by more than the minimum increment. The frequency of jump bids is 9–39% in the field and 40–61% in the lab. Jumping one's own bid is rare in the field (0%) and in the lab (0.09–0.16%). Thus, the magnitudes of these two simple statistics are suggestive of modest comparability. They also run probit regressions of what variables predict jump bids and find substantial overlap in sign and significance (their Table 9). In only one case is there a significant effect in the field–less jump bidding early in the auctions—which is significant and opposite in sign in the lab.

"Deal or No Deal"

Post et al. (2008) started with field data from the "Deal or No Deal" television show in three countries (United States, Germany, Holland). After making many simplifying assumptions, they could infer risky choice preference parameters from the contestants' decisions. The central interesting feature of their analysis is that decisions are consistent with risk tastes depending on prior outcomes in a sequence of choices. Specifically, contestants exhibit a "break-even effect" (taking risks to reach a previous point of reference) and a "house money" effect (taking more risks when the set of possible winning amounts shifts upward). They also conducted parallel lab experiments which had many features of the TV game show, including a grinning "game show host" (a popular lecturer at Erasmus University), a live audience, video cameras, and a computerized display of unopened briefcases, remaining prizes, and "bank offers." They report that "the audience was very excited and enthusiastic during the experiment, applauding and shouting hints, and most contestants showed clear symptoms of distress."

Post et al. (2008) concluded that

> Choices in the experiment are remarkably similar to those in the original TV show, despite the fact that the experimental stakes are only a small fraction of the original stakes. Consistent with the TV version, the break-even effect and the house-money effect also emerge in the experiments. (Post et al., 2008, p. 68)

The Weakest Link

In the game show "The Weakest Link," people take turns answering trivia questions, and participants can exclude others based on previous performance. Antonovics et al. (2009) compared field data from the TV show with a similar lab experiment, to see whether opponent gender makes a difference. They find that, in general, men do better against female opponents. However, the effect disappears in younger contestants. The effect is also absent in younger college students in a lab environment with higher stakes. Thus, going from a particular field setting to

a comparable lab setting yields the same results, provided that age and motivation are controlled for.

Dictator Sharing

Franzen and Pointner (2012) contrasted dictator-game sharing by students with the same students' responses to getting a "misdirected letter" which they could keep (pocketing 10 euros) or forward to its intended receiver. Two different studies varied the time gap between the lab experiment and the field experiment, either 4–5 weeks or 2 years !. They find modest correlations in pro-sociality between lab and field, around $\phi = 0.20$.

Lab–Field Generalization with Substantial Design and Subject Differences

A more common type of lab–field generalization correlates individual lab-based measurements of behavior or preferences with naturally occurring measures of socioeconomic activity from the same or different individuals (sometimes aggregated at the professional, community, firm, or national level). These analyses are not guaranteed to produce correlation, even if there is no fundamental lab–field difference ceteris paribus, because there are confounding differences in the lab and field settings. However, if there is some correlation in behavior, this is encouraging evidence that lab measures are associated with field behavior, even despite uncontrolled differences in lab and field.

For example, Barr and Serneels (2009) did trust experiments with Ghanian manufacturing workers, whose characteristics, output, and earnings were measured in a typical survey. In the experiments, workers generally repaid 1, 1.5, or 2 times as much as was invested by first-movers. Define those repaying more than 1.5 as "high reciprocators." They find that across firms, the proportion of high reciprocators is strongly correlated ($r = 0.55$, $p < 0.01$) with output per worker. High-reciprocator workers also earn 40% more ($t = 2.7$). However, instrumental control for reverse causality from earnings to reciprocity weakens the effect to an insignificant 17% increase ($t = 0.28$).

The results from recent studies of this type do typically find positive associations between lab and field measures. I briefly mention findings from a few studies organized by domain. My characterization of the studies' findings is meant to be not too aggressive about positive results: for example, many are statistically significant, but small in magnitude. More formal analyses should therefore certainly be conducted.

Pricing

- Comparable price elasticities (–0.69 and –0.79) in door-to-door field and lab sales of strawberries (Brookshire et al., 1987).

- Actual market share of "antibiotic-friendly" pork was not significantly different than a forecast based on a choice experiment outside a store in Oklahoma (Lusk et al., 2006).

Risk and Time Preference

- Impatience and risk tolerance are correlated with contribution to pensions by the self-employed in Chile (Barr and Packard, 2000).
- On a TV game show a simple game involving one or two spins of a wheel. In the field data and in a lab replica, subjects fail to take a second spin frequently when they should spin, in both field and lab (Tenorio and Cason, 2002).
- Bidders in a lab "Price is Right" replica exhibit two anomalies which are also observed in field data from the TV game show (Healy and Noussair, 2003): The last bidder does not optimize 36% of the time in the lab versus 43% in the field; and bids decline (as predicted by theory) only 7.7% of the time in the lab versus 12.1% in the field.
- Borrowers who are present-biased experimentally have larger credit card balances (Meier and Sprenger, 2008).
- Individual measures of time discounting from experimental amount-delay choices correlate modestly but reliably with various field data on self control (diet, exercise, saving, gambling, etc.) (Chabris et al., 2008). Correlations improve with aggregation.
- Female Connecticut trick-or-treaters who make an ambiguity-averse choice of a candy bag are more likely to wear a "less risky" (more popular) Halloween costume (Anagol et al., 2010).
- In price-list tests of risk aversion and time discounting in artificial lab experiments, using both student groups and citizen volunteers from a Danish representative sample, Andersen et al. (2010) "find no significant difference in the average degree of risk aversion and discount rates between the field [citizen] and laboratory samples." They also note that there is a common experimenter effect on discount rates in the two samples and suggest that it might be a "priming" effect in which a slower experimenter induced more patient choices.
- Slow adoption of modified Bt cotton by Chinese farmers is correlated with riska version and loss aversion, and well as lower overweighting of low probability (Liu, 2013).

Peer Effects

- A numerical estimate of the strength of peer effects on output in laboratory envelope-stuffing is "very similar to a comparable estimate derived by Ichino and Maggi (2000) with observational data" (Falk and Ichino, 2006).

Tournaments

- Subjects in lab effort-based tournaments, as well as fisherman in field competitions to catch the most fish, exert less effort when there are more competitors (as theory generally predicts) (List et al., 2010).

Pro-sociality

- Peruvian villagers who were less trustworthy players in trust experiments also defaulted on microfinance loans at higher rates (Karlan, 2005). However, the link between trust and loan repayment is less clear because trust can reflect altruistic giving, or an expectation of repayment. Karlan concludes (2005, p. 1698): "This endorses experimental economics as a valid measurement tool for field research, and the Trust Game as a valid method to measure trustworthiness, but **not** as a method to measure trust."
- Conditional cooperation and verbal disapproval in a public goods game predicts group fishing productivity in Japan (Carpenter and Seki, 2005).
- Giving in an experimental dictator game with charity recipients predicts whether Vermonters volunteer to fight fires (Carpenter and Myers, 2007).
- Experimental public goods contributions and patience for water predict limited (pro-social) common pool resource extraction among Brazilians who catch fish and shrimp (Fehr and Leibbrandt, 2008).
- An experimental bribery game done with Oxford students from around the world (in both an original sample and a replication) found that experimental bribery and acceptance among undergraduates, but not graduate students, was correlated with an index of corruption (Transparency International) in a student's home country (Barr and Serra, 2010).
- Experimental trustworthiness among University of Chicago MBAs correlates with their individual donations to a class gift (Baran et al., 2010).
- Dictator game allocations from Ugandan teachers to parents correlate with the teachers' actual teaching time (the inverse of absenteeism) (Barr and Zeitlin, 2010).
- Effort in a lab gift-exchange experiment with students in Lyon, France is positively correlated with both worker income and the worker's income rank in a comparison group (Clark et al., 2010). The same correlation signs and statistical strengths were also observed in ISSP survey questions from 17 OECD countries about willingness to "work harder than I have to in order to help the firm or organization I work for to succeed," as correlated with reported income and income rank.
- Relative favoritism of a low-status outgroup (the Khmer) by Vietnamese and Chinese in sharing and cooperation decisions is consistent with government policies favoring the Khmer (e.g., education and tax subsidies)

(Tanaka and Camerer, 2010) (no data are available at the individual level on policy preferences of the lab subjects, however).

- Left-handedness among women is correlated with selfishness in a lab dictator game and with measures of charitable donation in field surveys (Buser, 2010). Similarly, men are more trusting and reciprocal in a lab trust game and endorse the same behaviors in a LISS survey. (Some other lab behaviors and field measures do not show much association, however.)
- Group-level conditional cooperation in experiments is correlated with success in managing forest commons in Ethiopia (Rustagi et al., 2010).
- Experimental public goods (PG) contributions in 16 Indian villages correlate with how much salt each villager took from a common pool when he or she arrived to collect experimental earnings (Lamba and Mace, 2011). The individual-level correlations between PG contribution and salt-taking are low ($r = .057$), but the village-level correlations are extremely high ($r = 0.871$).[25] This study is an important empirical reminder that the level at which lab–field generalizability is best is not necessarily the individual level.
- Laboratory measures of pro-social behavior among truckers is correlated field pro-social behavior under comparable conditions (anonymous unrepeated interactions) (Anderson et al., 2011).
- Choices by Ethiopian nurses and doctors to work for NGOs correlated with their experimental pro-sociality in a generalized trust game (where a Responder can share money that one Proposer invested with a different Proposer). The correlation exists even after controlling for self-reported desire to help the poor (Serra et al., 2011).
- A low-income Dallas population who contribute more in public goods experiments are more pro-social in other experiments, and they self-report more donation and volunteering outside the lab (de Oliveira et al., 2011).
- Students making choices online about donating mosquito nets to pregnant mothers in Kenya from lottery winnings as well as donating to another anonymous student, make choices which are correlated across the two types of donees (Coffman, 2011).

Two studies show lab and field behavior that go in opposite directions:

- In theory and in lab experiments, all-pay auctions yield more revenue than public or anonymous voluntary contribution mechanisms (e.g., Schram and Onderstal, 2009). However, the opposite pattern was found (all-pay auctions yield the least) in a large door-to-door fundraising experiment with 4500 households (Onderstal et al., 2014). Note, even further, that in their field experiment Onderstal et al. (2014) found that VCM raised more revenue than lotteries, the opposite result of a field experiment by Landry et al. (2006). This observation is a simple reminder that field–field generalizability does not always work either: While the Onderstal et al. lab and field results do not match up well, neither do the results of the two different field experiments.[26]

- Results of corruption experiments conducted in Australia, India, Indonesia, and Singapore do not strongly correlate with national corruption indices (Cameron et al., 2009). Lab and field corruption are both high in India and low in Australia, but the lab and field results are mismatched in Indonesia and Singapore. The authors suggest that recent *changes* in public attitudes toward corruption could account for the two mismatches.

Several papers compare general regularities derived from field data (typically about the response to institutional changes) with highly stylized lab equivalents. In these cases there is a clear hope for generalizability in conducting the lab experiments, but the lab–field subject task and identity are not closely matched. However, the general finding is that comparative statics responses to institutional changes, as well as general regularities, often are similar in sign and magnitude in the field.

Kagel and Levin (1986) show correspondence to many properties of lab common-value auctions, showing overbidding and a "winner's curse," and they also report patterns in drainage leases in the Gulf of Mexico. Kagel and Roth (2000) created experimental matching markets with different features that are motivated by differences in observed behavior in naturally occurring markets, and they found parallel behavior in the lab and field. Blecherman and Camerer (1996) observed parallels between overbidding for baseball free agents (using field data) and overbidding in highly stylized lab experiments with features of free agency bidding. Bolton et al. (2013) studied changes in online auction reputation systems in simplified lab settings and showed that they are similar to field responses.

Conclusion: Where Do We Go from Here?

This chapter considers the issue of whether economic lab experiments should be expected to generalize to specific naturally occurring field settings. I suggest that generalizability of lab results is an exaggerated concern among non-experimenters for three possible reasons.

First, the scientific perspective that governed experimental economics from the beginning (Smith, 1976, 1982) is that all empirical methods are trying to accumulate regularity about how behavior is *generally* influenced by individual characteristics, incentives, endowments, rules, norms, and other factors. A typical experiment therefore has no specific target for "external validity"; the "target" is the general theory linking economic factors to behavior. (That's also the same "target" a typical field study has.) A special concern for external validity is certainly appropriate when the only goal of an experiment is to provide guidance about how behavior might work in a specific external setting—in Campbell and Stanley's "known character" education language. But such targeted guidance is rarely the goal of experiments in economics.

Second, when experiments are criticized for limited generalizability (as by LL in many passages), that criticism depends on contrasting stereotypes of a canonical low-stakes, artificial experiment with students and a canonical field setting with self-selected skilled agents and high stakes. Criticisms that depend on these contrasted stereotypes ignore the crucial fact that experiments can be very different and that *more experiments can always be conducted*. Since many different types of experiments can be run, the threats to external validity that LL note–moral considerations, obtrusiveness, context, self-selection, and stakes—can typically be varied in experiments to see how much they matter (as Falk and Heckman (2009) noted too).

Third, non-experimenters in economics often do not realize the extent to which modern economics experiments, which developed in the 1970s, differ very sharply from methods in older stereotypical psychology experiments. Pioneering experimental economists developed their own procedures and techniques specifically to induce careful control over preference and subject motivation, in a way that is largely absent in experimental psychology. However, LL quote sources from the 1960s asserting that the experimenter–subject relationship is "matched only by that of . . . drill sergeant and trainee" or that subjects are "scientific do-gooders." It is certainly true that in a stereotypical *psychology* experiment, there is often no clear performance metric and no incentive pay (students often must participate for "course credit," perhaps reluctantly), there is sometimes deception, and fears about demand effects are reasonable since the experimenters often really do want the experiment to "work" (i.e., give a particular outcome). But these features are *not* typical of economics experiments. We view participants as students looking for flexible part-time work that might be more interesting than comparable spot-market alternatives (as Slonim et al. (2013) and Abeler and Nosenzo (in press) document).

Let's end by reflecting on one ominous warning in Levitt and List's extensive writing about lab-generalizability:

> Yet unless considerable changes are made in the manner in which we conduct lab experiments, our model highlights that the relevant factors will rarely converge across the lab and many field settings. (LL, 2007a, p. 364)

Fortunately, their warning of rare convergence seems to be wrong about the most direct lab–field comparisons. It is also wrong for most indirect lab–field comparisons, including those generalizing pro-sociality from the lab to the field.

NOTES

This chapter was prepared for an NYU Methods conference and edited book. Citations from many ESA members were extremely helpful. Support from HFSP and the Gordon and Betty Moore Foundation are gratefully acknowledged. Thanks to Dan Knoepfle, Stephanie Wang, Alec Smith, and especially Rahul Bhui for research assistance, to many

patient listeners, to Tim Salmon and Abigail Barr for candid feedback, and to John List for his JPE (List, 2006) data and discussions.

1. There is also a lively debate about the value of instrumental variable IV studies versus other techniques (e.g., Heckman and Urzua, 2010; Imbens, 2009; Deaton 2009).

2. Uri Gneezy made this point in a panel discussion at a Berkeley SEED-EC conference, December 2, 2011.

3. See Banerjee and Duflo (2009).

4. Chris Udry made this point in a panel discussion at a Berkeley SEED-EC conference, December 2, 2011.

5. On subject pool effects, see Ball and Cech (1996), Fréchette (2015), and Henrich et al. (2010).

6. This figure comes from Cherry et al.'s (2002) $10 EL condition, with stakes comparable to many other games, but some other giving rates in their earned-income (E) conditions were lower and higher, depending on stakes and experimenter-blindness.

7. Camerer and Thaler (1995, p. 216) predicted the earned–unearned difference in dictator giving shown by Cherry et al. (2002), writing: "Etiquette may require you to share a windfall with a friend, but it certainly does not require you to give up some of your hard-earned year-end bonus to a stranger."

8. See also Andreoni and Bernheim (2009), Krupka and Weber (2008), Bardsley (2005), and List (2007).

9. Others (e.g. LL, 2008) use the term "scrutiny," which will be treated as synonymous with the more established term "obstrusiveness" in this chapter.

10. Many field experiments, in development for example, are clearly as obtrusive as in university labs and may create even stronger potential experimenter effects if the subject–experiment relationship is less familiar and influential to low-education rural villagers, say, than it is to college students. Cilliers, Dube and Siddiqi (2013) report an example from an artificial field experiment, showing that presence of white foreigners increases dictator donations in Sierra Leone.

11. An important example is the double-blind dictator experiments (Hoffman et al., 1998). One possible interpretation of those results is that subjects in the standard game give because they believe that the experimenter wants them to. (Note that in this design, self-interested behavior also occurs at a boundary of the choice set, so any random deviation looks like both pro-social giving and like expression of satisfying a perceived demand to give.) The drop in giving in the double-blind case could be consistent with disappearance of such a perceived demand to give. But it could also be consistent with changing the subject's belief that their specific allocation will be known to anybody, not just to an experimenter.

12. Experimental economists work hard to eliminate demand effects. Orne (1962) advocates trying to measure demand effects in various ways (e.g., presenting passive subjects with the experimental materials and asking them what they think is likely to happen or what the experimenter expects). More studies using innovative methods would be useful. For example, Bischoff and Frank (2011) tried to create a "social instructor demand effect" (cf. Zizzo, 2010) by hiring an (incentivized) actor to read instructions.

13. They also included a truly transparent new experiment (whether a person was more likely to buy a $10 school raffle ticket if they were given a free can of soda), and 88% of the subjects did correctly guess the experimenters' hypothesis.

14. After figuring out how a public goods game worked, an Orma participant in Kenya exclaimed "Harrambee!" because the game reminded her of a practice used to raise funds for public goods (such as building a school) in which a village official harangues people to contribute one by one (Ensminger, 2004).

15. Selection is complicated. Overconfidence, discrimination, and social network effects all mitigate the idea that there is pure self-selection of people into markets those people are best suited for.

16. *Note*: All descriptions of their results are drawn from their Table 4.7, which regresses participation on all individual characteristics. In that specification, a fifth pro-sociality measure—giving to charity in a specific experimental decision—was associated with participation ($p < 0.10$), but this is a weak effect given the large sample size N. In a specification of participation using only the five volunteering variables, the pseudo-R^2 is .054.

17. Al-Ubaydli and List (2015) note that "the covertness implicit in a NFE, which we are arguing is desirable, is sometimes impossible . . .."

18. Andersen et al.'s (2011) study of high-stakes ultimatum game bargaining among the Khasi speakers in India does show an apparent decrease in rejections of percentage offers as stakes increase (though typical rejected offer amounts also increase with stakes). Their result actually supports the conclusion from most modest-stakes ultimatum experiments, which is that rejections express some type of social preference, because trading off monetary share and money, as in Bolton and Ockenfels (2000) for example, easily explains the observed stakes effect.

19. Data were generously supplied by John List. Al-Ubaydli and List (2015) note that "Camerer is one of many scholars who have asked for the List (2006) data. Each one of them has delivered replication in the first sense of the definition above [reproducing statistical results from original data]; the data show what List (2006) reported. In addition, all of them but Camerer have been satisfied with the conclusions drawn (or at least we have not heard of any dismay." First, it would be useful to know more details about the breadth and conclusions from these many scholars. Second, and more importantly, the "replication" here also includes *new* analysis of within-subject patterns which was not reported in List (2006).

20. Thanks to John Kagel for pointing this out.

21. Boyd and Richerson's explanation for exceptional human altruism is that genes and culture coevolved in humans to create preferences for pro-social behavior and supporting adaptations (such as punishment).

22. There is ongoing debate about how well mixed equilibrium describes behavior, under various conditions. Binmore et al. (2001) report good accuracy. Camerer (2003, Chapter 3) reports a high correlation (0.84) between equilibrium predicted frequencies and actual frequencies across a variety of games. Martin et al. (2014) report extreme accuracy with chimpanzee subjects.

23. Engelmann and Hollard (2010) essentially replicate this effect, by showing that forced trading reduces endowment effects (consistent with the interpretation that trading frequency mutes the effect of loss-aversion by shifting expectations of likely trade; see Della Vigna (2009, p. 328). Feng and Seasholes (2005) and Dhar and Zhu (2006) report similar effects of investor sophistication and experience on reduction in the financial disposition effect (the tendency to oversell winning stocks and undersell losing stocks).

24. The most widely cited are defined for this purpose as those with more than 100 Google Scholar cites, as of 24 April 2011.

25. Their paper also has excellent data on individual-level characteristics, which is important for drawing conclusions between ethnic groups and villages. See also Henrich et al. (2012) for a comment disputing some of Lamba and Mace's conclusions.

26. Thanks to Arthur Schram for pointing this out.

REFERENCES

Abeler, J. and F. Marklein. 2010. Fungibility, Labels and Consumption. IZA Discussion Paper No. 3500.

Abeler, J. and D. Nosenzo. In press. Self-Selection into Laboratory Experiments: Pro-social Motives vs. Monetary Incentives. *Experimental Economics*.

Alatas, V., L. Cameron, A. Chaudhuri, N. Erkal, and L. Gangadharan. 2009. Subject Pool Effects in a Corruption Experiment: A Comparison of Indonesian Public Servants and Indonesian Students. *Experimental Economics* **12**(1):113–132.

Allcott, H. and S. Mullainathan. 2012. External Validity and Partner Selection Bias. NBER Working Paper No. 18373.

Alpizar, F., Carlsson, F., and Johansson-Stenman, O. 2008. Does Context Matter More for Hypothetical than for Actual Contributions? Evidence from a Natural Field Experiment. *Experimental Economics* **11**(3):299–314.

Al-Ubaydli, O. and J. List. 2015. On the Generalizability of Experimental Results in Economics. In *Handbook of Experimental Economic Methodology,* eds. Fréchette, G. R. and A. Schotter. Oxford University Press.

Anagol, S., S. Bennett, G. Bryan, T. Davenport, N. Hite, D. Karlan, P. Lagunes, and M. McConnell. 2010. There's Something About Ambiguity. Working paper.

Andersen, S., S. Ertac, U. Gneezy, M. Hoffman, and J. A. List. 2011. Stakes Matter in Ultimatum Games. *American Economic Review.* **101**:3427–3439

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2008. Risk Aversion in Game Shows. In *Risk Aversion in Experiments: Research in Experimental Economics, Volume* 12, eds. G. W. Harrison and J. Cox. Emerald Group Publishing/JAI Press, pp. 359–404.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2010. Preference Heterogeneity in Experiments: Comparing the Field and Laboratory. *Journal of Economic Behavior and Organization* **73**(2):209–224.

Anderson, J., M. Bombyk, S. Burks, J. Carpenter, D. Ganzhorn, L. Goette, D. Nosenzo, and A. Rustichini. 2011. Lab Measures of Other-Regarding Behavior Predict Some Choices in a Natural On-the-Job Social Dilemma: Evidence from Truckers. Working paper.

Anderson, J., S. Burks, J. Carpenter, L. Goette, K. Maurer, D. Nosenzo, R. Potter, K. Rocha, and A. 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**:170–189.

Andreoni, J. and B. Bernheim. 2009. Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica* **77**(5):1607–1636.

Angrist, J. and A. Krueger. 2001. Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* **15**(4):69–85.

Antonovics, K., P. Arcidiacono, and R. Walsh. 2009. The Effects of Gender Interactions in the Lab and in the Field. *Review of Economics and Statistics* **91**(1):152–162.

Armentier, O. and A. Boly. 2013. Corruption in the Lab and in the Field in Burkina Faso and in Canada. *Economic Journal* **123**(573):1168–1187.

Ball, S. B. and P.-A. Cech. 1996. Subject Pool Choice and Treatment Effects in Economic Laboratory Research. In *Research in Experimental Economics, Volume 6*, ed. R. Mark Isaac. Amsterdam: Elsevier Science and Technology Books, pp. 239–292.

Bandiera, O., I. Barankay, and I. Rasul. 2005. Social Preferences and the Response to Incentives: Evidence from Personnel Data. *Quarterly Journal of Economics* **120**(3):917–962.

Banerjee, A. and E. Duflo. 2009. The Experimental Approach to Development Economics. *Annual Review of Economics* **1**:151–178.

Baran, N. M., P. Sapienza, and L. Zingales. 2010. Can We Infer Social Preferences from the Lab? Evidence from the Trust Game. NBER Working Paper.

Bardsley, N. 2005. Experimental Economics and the Artificiality of Alteration. *Journal of Economic Methodology* **12**:239–251.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden. 2009. *Experimental Economics: Rethinking the Rules*. Princeton, NJ: Princeton University Press.

Barr, A. and T. Packard. 2000. Revealed and Concealed Preferences in the Chilean Pension System: An Experimental Investigation. Department of Economics Discussion Paper Series, University of Oxford.

Barr, A. and P. Serneels. 2009. Reciprocity in the Workplace. *Experimental Economics* **12**(1):99–112.

Barr, A. and D. Serra. 2010. Corruption and Culture: An Experimental Analysis. *Journal of Public Economics* **94**(11–12):862–869.

Barr, A. and A. Zeitlin. 2010. Dictator Games in the Lab and in Nature: External Validity Tested and Investigated in Ugandan Primary Schools. CSAE Working Paper Series, Centre for the Study of African Economies, University of Oxford.

Bateman, I., A. Munro, B. Rhodes, C. Starmer, and R. Sugden. 1997. A Test of the Theory of Reference-Dependent Preferences. *Quarterly Journal of Economics* **112**(2):479–505.

Belot, M., R. Duch, and L. Miller. 2010. Who Should be Called to the Lab? A Comprehensive Comparison of Students and Non-students in Classic Experimental Games. Nuffield Centre for Experimental Social Sciences, University of Oxford, Discussion Paper Series.

Benz, M. and S. Meier. 2008. Do People Behave in Experiments as in the Field? Evidence from Donations. *Experimental Economics* **11**(3):268–281.

Binmore, K., J. Swierzbinski, and C. Proulx. 2001. Does Minimax Work? An Experimental Study. *Economic Journal* **111**:445–465.

Bischoff, I. and B. Frank. 2011. Good news for experimenters: Subjects are Hard to Influence by instructors' Cues. *Economics Bulletin* **31**(4):3221–3225.

Blecherman, B. and C. F. Camerer. 1996. Is There a Winner's Curse in the Market for Baseball Players? Evidence from the Field. Social Science Working Paper, California Institute of Technology.

Bolton, G., B. Greiner, and A. Ockenfels. 2013. Engineering Trust—Reciprocity in the Production of Reputation Information. *Management Science* **59**(2):265–285.

Bolton, G. E. and A. Ockenfels. 2000. ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* **90**(1):166–193.

Boyd, R. and P. J. Richerson. 2009. Culture and the Evolution of Human Cooperation. *Philosophical Transactions of the Royal Society B* **364**:3281–3288.

Bracht, G. H. and G. V. Glass. 1968. The External Validity of Experiments. *American Educational Research Journal* **5**:437–474.

Brookshire, D. S, D. L. Coursey, and W. D. Schulze. 1987. The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior. *Economic Inquiry* **25**(2):239–250.

Buser, T. 2010. Handedness Predicts Social Preferences: Evidence Connecting the Lab to the Field. Tinbergen Institute Paper TI 2010-119/3.
Camerer, C. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton, NJ: Princeton University Press.
Camerer, C., T.-H. Ho, and J.-K. Chong. 2004. A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics* **119**(3):861–898.
Camerer, C. and R. Hogarth. 1999. The Effects of Financial Incentives in Experiments: A Review and Capital–Labor–Production Framework. *Journal of Risk and Uncertainty* **19**(1–3):7–42.
Camerer, C. and R. Thaler. 1995. Anomalies: Ultimatums, Dictators and Manners. *Journal of Economic Perspectives* **9**(2):209–219.
Cameron, L., A. Chaudhuri, N. Erkal, and L. Gangadharan. 2009. Propensities to Engage in and Punish Corrupt Behavior: Experimental Evidence from Australia, India, Indonesia and Singapore. *Journal of Public Economics* **93**(7–8):843–851.
Campbell, D. and J. Stanley. 1963. Experimental and Quasi-Experimental Designs for Research on Teaching. In *Handbook of Research on Teaching*, ed. N. L. Gage. Chicago: Rand McNally.
Carlsson, F., H. He, and P. Martinsson. 2013. Easy Come, Easy Go–The Role of Windfall Money in Lab and Field Experiments. *Experimental Economics* **16**:190–207.
Carpenter, J. and C. K. Myers. 2007. Why Volunteer? Evidence on the Role of Altruism, Reputation, and Incentives. IZA Discussion Papers.
Carpenter, J. and E. Seki. 2005. Competitive Work Environments and Social Preferences: Field Experimental Evidence from a Japanese Fishing Community. Middlebury College Working Paper Series.
Chabris, C. F., D. I. Laibson, C. L. Morris, J. P. Schuldt, and D. Taubinsky. 2008. Individual Laboratory-measured Discount Rates Predict Field Behavior. *Journal of Risk and Uncertainty* **37**:237–269.
Cherry, T. L., P. Frykblom, and J. Shogren. 2002. Hardnose the Dictator. *American Economic Review* **92**(4):1218–1221.
Cilliers, J., Dube, O., Siddiqi, B. 2013. 'White Man's Burden?' A Field Experiment on Generosity and Foreigner Presence. University of Oxford working paper.
Clark, A., D. Masclet and M. C. Villeval. 2010. Effort and Comparison Income. Experimental and Survey Evidence. *Industrial and Labor Relations Review* **63**(3):407–426.
Cleave, B., N. Nikiforakis, and R. Slonim. 2012. Is There Selection Bias in Laboratory Experiments? The Case of Social and Risk Preferences. *Experimental Economics* **16**(3), 372–382.
Clement, D. 2002. Interview with Gary Becker. In *The Region*. The Federal Reserve Bank of Minneapolis.
Coffman, L. C. 2011. Intermediation Reduces Punishment (and Reward). *American Economic Journal: Microeconomics.* **3**:77–106.
Deaton, A. 2009. Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. NBER Working Paper.
Della Vigna, S. 2009. Psychology and Economics: Evidence from the Field. *Journal of Economic Literature* **47**:315–372.
de Oliveira, A., R. Croson, and C. C. Eckel. 2011. The Giving Type: Identifying Donors. *Journal of Public Economics* **95**(5–6):428–435.

Dhar, R. and N. Zhu. 2006. Up Close and Personal: Investor Sophistication and the Disposition Effect. *Management Science* **52**(5):726–740.

Eckel, C. C and P. J Grossman. 1996. Altruism in Anonymous Dictator Games. *Games and Economic Behavior* **16**:181–191.

Eckel, C. C. and P. J. Grossman. 2000. Volunteers and Pseudo-Volunteers: The Effect of Recruitment Method in Dictator Experiments. *Experimental Economics* **3**(2):107–120.

Engelmann, D. and G. Hollard. 2010. Reconsidering the Effect of Market Experience on the "Endowment Effect". *Econometrica* **78**(6):2005–2019.

Ensminger, J. E. 2004. Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* eds. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. London: Oxford University Press.

Falk, A. and J. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* **326**(5952):535–538.

Falk, A. and A. Ichino. 2006. Clean Evidence on Peer Effects. *Journal of Labor Economics* **24**(1):39–57.

Falk, A., Meier, S. and C. Zehnder. 2013. Do Lab Experiments Misrepresent Social Preferences? The Case of Self-selected Student Samples. *Journal of the European Economic Association* **11**(4): 839–852.

Fehr, E. and A. Leibbrandt. 2008. Cooperativeness and Impatience in the Tragedy of the Commons. *Journal of Public Economics* **95**:1144–1155.

Feng, L. and M. Seasholes. 2005. Do Investor Sophistication and Trading Experience Eliminate Behavioral Biases in Financial Markets? *Review of Finance* **9**(3):305–351.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M., 1994. Fairness in Simple Bargaining Experiments. *Games and Economic Behavior* **6**:347–369.

Franzen, A. and S. Pointner. 2012. The External Validity of Giving in the Dictator Game: A Field Experiment Using the Misdirected Letter Technique. *Experimental Economics*.

Fréchette, G. R. 2015. Laboratory Experiments: Professionals Versus Students. In *Handbook of Experimental Economics Methodology*, eds. Guillaume Fréchette and Andrew Schotter. New York: Oxford University Press, Chapter 17, present volume.

Gelman, Andrew. 2010. Another Update on the Spam email Study, Blog post http://andrewgelman.com/2010/05/anotherupdate/.

Haigh, M. and J. A. List. 2005. Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis. *Journal of Finance* **60**(1):523–534.

Healy, P. J. and C. Noussair. 2004. Bidding Behavior in the Price Is Right Game: An Experimental Study. *Journal of Economic Behavior and Organization* **54**(2):231–247.

Heckman, J. and S. Urzua. 2010. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. *Journal of Econometrics* **156**(1):27–37.

Henrich, J., R. Boyd, S. Bowles, C. F. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, et al. 2005. "'Economic Man' in Cross-Cultural" Perspective: Behavioral Experiments in 15 Small-Scale Societies. *Behavioral and Brain Sciences* **28**(6):795–815.

Henrich, J., R. Boyd, R. McElreath, M. Gurven, P. J. Richerson, J. Ensminger, M. Alvard, A. Barr, C. Barrett, A. Bolyanatz, C. Camerer, J. C. Cardenas, E. Fehr, H. Gintis, F. Gil-White, E. Gwako, N. Henrich, K. Hill, C. Lesorogol, J. Q. Patton, F. Marlowe, D. Tracer, and J. Ziker. In 2012. Culture Does Account for Variation in Game Behaviour. *Proceedings of the National Academy of Sciences*. **109**(2):E32–E33

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, J. Ziker. 2006. Costly Punishment Across Human Societies. *Science* **312**:1767–1770.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. The Weirdest People in the World? *Behavioral and Brain Sciences* **33**:61–83.

Hoffman, E., K. A. McCabe, and V. L. Smith. 1998. Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology. *Economic Inquiry* **36**(3):335–352.

Hoffman, M. H. and J. Morgan. 2011. Who's Naughty? Who's Nice? Social Preferences in Online Industries. Working paper.

Ichino, A. and G. Maggi. 2000. Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm. *Quarterly Journal of Economics* **115**(3):1057–1090.

Imbens, G. 2009. Better Late Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). NBER Working Paper.

Isaac, R. M. and K. Schnier. 2006. Sealed Bid Variations on the Silent Auction, In *Experiments Investigating Fundraising and Charitable Contributors (Research in Experimental Economics,* Volume 11), eds. R. M. Isaac and D. D. Davis. West Yorkshire, England: Emerald Group Publishing Limited, pp. 31–46.

Kagel, J. and D. Levin. 1986. The Winner's Curse and Public Information in Common Value Auctions. *American Economic Review* **76**:894–920.

Kagel, J. H. and A. E. Roth. 2000. The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment. *Quarterly Journal of Economics* **115**(1):201–235.

Karlan, D. S. 2005. Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *American Economic Review* **95**(5):1688–1699.

Krupka, E. and R. Weber. 2008. Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? IZA Discussion Paper Institute for the Study of Labor. No. 3860.

Lamba, S. and R. Mace. 2011. Demography and Ecology Drive Variation in Cooperation across Human Populations. *Proceedings of the National Academy of Sciences.* **108**(35):14426–14430

Lambdin, C. G. and V. A. Shaffer. 2009. Are Within-Subjects Designs Transparent? *Judgment and Decision Making* **4**(7):554–566.

Landry, C., A. Lange, J. List, M. Price, and N. Rupp. 2006. Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment. *Quarterly Journal of Economics* **121**(2):747–782.

Lazear, E. P., U. Malmendier, and R. A. Weber. 2012. Sorting in Experiments with Application to Social Preferences. *American Economic Journal: Applied Economics.* **4**(1):136–163

Levitt, S. and S. Dubner. 2009. *Super Freakonomics*. New York: HarperCollins.

Levitt, S. and J. A. List. 2007a. Viewpoint: On the Generalizability of Lab Behaviour to the Field. *Canadian Journal of Economics* **40**(2):347–370.

Levitt, S. and J. A. List. 2007b. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World. *Journal of Economic Perspectives* **21**(2):153–174.

Levitt, S. and J. A. List. 2008. Homo Economicus Evolves. *Science* **319**(5865):909–910.

Levitt, S. D., J. A. List, and D. Reiley. 2010. What Happens in the Field Stays in the Field: Professionals Do Not Play Minimax in Laboratory Experiments. *Econometrica* **78**(4):1413–1434.

List, J. A. 2002. Preference Reversals of a Different Kind: The More Is Less Phenomenon. *American Economic Review* **92**(5):1636–1643.

List, J. A. 2003. Does Market Experience Eliminate Market Anomalies? *Quarterly Journal of Economics* **118**(1):41–71.

List, J. A. 2004. Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace. *Econometrica* **72**(2):615–625.

List, J. A. 2006. The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy* **114**(1):1–37.

List, J. A. 2007. On the Interpretation of Giving in Dictator Games. *Journal of Political Economy* **115**(3):482–494.

List, J. A. 2009. The Economics of Open Air Markets. NBER Working Paper.

List, J. A., D. van Soest, J. Stoop, and H. Zhou. 2010. Optimal Contest Design When Common Shocks Are Skewed: Theory and Evidence from Lab and Field Experiments. http://www:mcgill.ca/files/economics/Tournamentshock20100312.pdf

Liu, E. M. 2013. Time to Change What to Sow: Risk Preferences and Technology Adoption Decisions of Cotton Farmers in China. *Review of Economics and Statistics*. **95**(4):1386–1403.

Lusk, J. L., F. B. Norwood, and J. R. Pruitt. 2006. Consumer Demand for a Ban on Antibiotic Drug Use in Pork Production. *American Journal of Agricultural Economics* **88**(4):1015–1033.

Martin, C., R. Bhui, P. Bossaerts, T. Matsuzawa, and C. Camerer, et al. 2012. Chimpanzee Choice Rates in Competitive Games Match Equilibrium Game Theory Predictions. *Nature Scientific Reports* **4**(5182), June 2014.

Meier, S. and C. Sprenger. 2010. Present-Biased Preferences and Credit Card Borrowing. *American Economic Journal: Applied Economics* **2**(1):193–210.

Milkman, K.L., M. Akinola, and D. Chugh. 2012. Temporal Distance and Discrimination: An Audit Study in Academia. *Psychological Science* **27**:710–717.

Onderstal, S., A. Schram, and A. Soetevent. 2014. Reprint of: Bidding to Give in the Field *Journal of Public Economics* **114**:87–100.

Orne, M. T. 1962. On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications. *American Psychologist* **17**(11):776–783.

Östling, R., J. T. Wang, E. Y. Chou, and C. F. Camerer. 2011. Testing Game Theory in the Field: Swedish LUPI Lottery Games. *American Economic Journal: Microeconomics* **3**(3):1–33.

Palacios-Huerta, I. 2003. Professionals Play Minimax. *Review of Economic Studies* **70**(2):395–415.

Palacios-Huerta, I. and O. Volij. 2008. Experientia Docet: Professionals Play Minimax in Laboratory Experiments. *Econometrica* **76**(1):71–115.

Pierce, A. H. 1908. The Subconscious Again. *Journal of Philosophy, Psychology and Scientific Methods* **5**(10):264–271.

Post, T., M. J. van den Assem, G. Baltussen, and R. H. Thaler. 2008. Deal or No Deal? Decision Making under Risk in a Large-Payoff Game Show. *American Economic Review* **98**(1):38–71.

Pruitt, D. and M. Kimmel. 1977. Twenty Years of Experimental Gaming: Critique, Synthesis, and Suggestions for the Future. *Annual Revue of Psychology* **28**:363–392.
Rapoport, A. 1970. Conflict Resolution in the Light of Game Theory and Beyond. In *The Structure of Conflict*, ed. P. Swingle. New York: Academic Press.
Rebitzer, J. B. 1988. Unemployment, Labor Relations and Unit Labor Costs. *American Economic Review* **78**(2):389–394.
Rustagi, D., S. Engel, and M. Kosfeld. 2010. Conditional Cooperation and Costly Monitoring Explain Success in Forest Commons Management. *Science* **330**(6006):961–965.
Samuelson, P. and W. Nordhaus. 1985. *Economics*. New York: McGraw-Hill.
Schram, A. and S. Onderstal. 2009. Bidding to Give: An Experimental Comparison of Auctions for Charity. *International Economic Review* **50**(2):431–457.
Schultz, D. P. 1969. The Human Subject in Psychological Research. *Psychological Bulletin* **72**(3):214–228.
Serra, D., P. Serneels, and A. Barr. 2011. Intrinsic Motivations and the Non-profit Health Sector: Evidence from Ethiopia. *Personality and Individual Differences* **51**(3):309–314.
Sims, C. 2010. But Economics Is Not an Experimental Science. *Journal of Economic Perspectives* **24**(2):59–68.
Slonim, R., Wang, C., Garbarino, E., and D. Merrett. 2013. Opting-in: Participation Bias in Economic Experiments. *Journal of Economic Behavior and Organization* **90**, 43–70.
Smith, V. L. 1976. Experimental Economics: Induced Value Theory. *American Economic Review* **66**(2):274–279.
Smith, V. 1982. Microeconomic Systems as an Experimental Science. *American Economic Review* **72**(5):923–955.
Smith, V. and J. Walker. 1993. Rewards, Experience, and Decision Costs in First Price Auctions. *Economic Inquiry* **31**(2):237–244.
Snowberg, E. C. and J. Wolfers. 2010. Explaining the Favorite-Longshot Bias: Is It Risk-Love or Misperceptions? *Journal of Political Economy* **118**(4):723–746.
Sonnemann, U., C. F. Camerer, C. R. Fox, and T. Langer. 2013. Partition Dependence in Prediction Markets: Field and Lab Evidence. *PNAS* **110**(29):11779–11784.
Stanley, O. February 22, 2011. Chicago Economist's 'Crazy Idea' Wins Ken Griffin's Backing. *Bloomberg Markets Magazine*. http://www.bloomberg.com/news/2011-02-23/chicago-economist-s-crazy-idea-for-education-wins-ken-griffin-s-backing.html.
Stoop, J. 2014. From the Lab to the Field: Envelopes, Dictators and Manners. *Experimental Economics* **17**:304–313.
Stoop, J., C. Noussair, and D. van Soest. 2012. From the Lab to the Field: Public Good Provision with Fishermen. *Journal of Political Economy* **120**(6):1027–1056.
Tanaka, T. and C. F. Camerer. 2010. Patronizing Economic Preferences toward Low-Status Groups in Vietnam. Working paper.
Tenorio, R. and T. Cason. 2002. To Spin or Not to Spin? Natural and Laboratory Experiments from The Price is Right. *Economic Journal* **112**:170–195.
Webb, E., D. Campbell, R. Schwartz, and L. Sechrest. 1999. *Unobtrusive Measures: Revised Edition*. Thousand Oaks, CA: SAGE Publications, Inc.
Wooders, J. 2010. Does Experience Teach? Professionals and Minimax Play in the Lab. *Econometrica* **78**(3):1143–1154.
Zizzo, D. 2010. Experimenter Demand Effects in Economic Experiments. *Experimental Economics* **13**:75–98.

CHAPTER 15

THEORY, EXPERIMENTAL DESIGN, AND ECONOMETRICS ARE COMPLEMENTARY (AND SO ARE LAB AND FIELD EXPERIMENTS)

GLENN W. HARRISON, MORTEN I. LAU, AND E. ELISABET RUTSTRÖM

Introduction

Experiments are conducted with various purposes in mind including theory testing, mechanism design, and measurement of individual characteristics. In each case a careful researcher is constrained in the experimental design by prior considerations imposed by either theory, common sense, or past results. We argue that the integration of the design with these elements needs to be taken even further. We view all these elements that make up the body of research methodology in experimental economics as mutually dependent and therefore take a systematic

approach to the design of our experimental research program. Rather than drawing inferences from individual experiments or theories as if they were independent constructs, and then using the findings from one to attack the other, we recognize the need to constrain the inferences from one by the inferences from the other. Any data generated by an experiment needs to be interpreted jointly with considerations from theory, common sense, complementary data, econometric methods, and expected applications.

We illustrate this systematic approach by reference to a research program centered on large artifactual field experiments we have conducted in Denmark.[1] The motivation for our research was to generate measures of household and individual characteristics for use in a range of policy valuations. An important contribution that grew out of our work is the complementarity of lab and field experiments.

One such characteristic was the risk preferences of representative Danish residents. Predicted welfare effects from policy changes are always uncertain, in part because of imprecisely known parameter values in the policy simulation models used. We introduce the term "policy lottery" to refer to such uncertainties over the predicted policy effects. In light of these uncertainties, we argue that the welfare impact calculated for various households should reflect their risk attitudes. When comparing policies with similar expected benefits but with differences in the uncertainty over those predicted effects, a risk-averse household would prefer the policy with less uncertain effects to that with more. Including measures of risk attitudes in policy evaluations can therefore have important implications for inferences about the distribution of welfare effects. This is a significant improvement over the standard practice in policy evaluations that either assume risk neutrality or some arbitrarily selected risk coefficient employed uniformly over all household types. Our dominating justification for the expense of going out in the field derived from the policy need to provide measures for households and individuals that are representative of the general Danish population.

An instrumental part of our research program was the inclusion of a number of complementary lab experiments conducted at a much lower cost because of the use of convenience subject pools: students. Due to the lower cost, we could conduct a wider range of robustness tests varying elicitation instruments and procedures; but because we sampled from a more restricted population, these results are not by themselves informative regarding the policy applications we have in mind. Nevertheless, the results obtained from such convenience samples can be used to condition the inferences drawn from the observations on the field sample.

Another aspect of the systematic approach was to use several theoretical considerations to guide our experimental design from the start. One important characteristic that we measure is the discount rate of individuals across various household types. Theory is quite clear that what is being discounted is not the money stream but the stream of utility that derives from that money. Recognition of this fact had an influence not only on the inclusion of tasks incorporating both risk and time manipulations but also on the econometric strategy of joint

estimation. The joint estimation approach leads to estimates of risk attitudes that are consistent with the estimated discount rates and vice versa.

Finally, the systematic approach we advocate encourages the use of common sense constraints on the inferences drawn from the data. For example, many structural model specifications suffer from inflexibility globally so that they provide poor predictions on domains outside the one on which the data was generated. The Constant Relative Risk Aversion function, for example, if estimated on small stakes, can make predictions on large stakes that may appear ridiculous. The same may even be the case for the more flexible Expo-power function if estimated on a stake domain where the income effect is negligible. Inferences drawn from estimations using restrictive domains and restrictive specifications must therefore be constrained with common sense constraints on their applicability.

In the section entitled "Policy Lotteries," we introduce the concept of policy lotteries, giving a few examples. In the section entitled "Risk Aversion," we discuss how we draw inferences about risk attitudes using our systematic approach that includes conditioning these inferences on smaller-scale lab experiments, on sample selection effects and elicitation methods, on econometric and statistical strategies such as sampling frame and structural estimation approaches, and on theoretical and common sense considerations about out-of-domain predictions. In the section entitled "Discount Rates," we discuss inferences about discount rates and demonstrate the power of joint estimation of risk and time preferences as motivated by theory. The section entitled "Lessons Learned" expands the joint inference discussion to longitudinal issues such as temporal stability.

Policy Lotteries

The motivation for the field experiments on which this research program is centered came from our earlier work with the Danish Ministry of Business and Industry between 1996 and 2000 to develop computable general equilibrium (CGE) models of public policy. Those policies ranged from general tax reforms to specific carbon tax reforms, from the effects of relaxing domestic retail opening hours to the effects on Denmark of global trade reform, and from intergenerational welfare issues to the dynamics of human capital formation. One of the hallmarks of the CGE models we were developing was an explicit recognition that many of the structural parameters of those models were uncertain and that policy recommendations that came from them amounted to a "policy lottery" in which probabilities could be attached to a range of possible outcomes. Recognition that the simulated effects of policy on households were uncertain, because the specific parameters of the model were uncertain, meant that a proper welfare analysis needed to account for the risk attitudes of those households.

Related to this dimension of these simulated results, in many cases there were nontrivial intertemporal trade-offs, such as foregone welfare in the short-term in return for longer-term gains. Indeed, this trade-off is a common feature of dynamic

CGE policy models (e.g., Harrison et al., 2000). Obviously the proper welfare evaluation needed to also account for the subjective discount rates that those households employed. For example, one of the policy issues of interest to the Danish government was why Danes appeared to "underinvest" in higher education. We elicited discount rates, in part, to address that policy question directly (see Lau (2000)).

A policy lottery is a representation of the predicted effects of a policy in which the uncertainty of the simulated impact is explicitly presented to the policy maker. Thus when the policy maker decides that one policy option is better than another, the uncertainty in the estimate of the impact has been taken into account. Note that this is uncertainty in the *estimate of the impact*, and not necessarily uncertainty in the *impact itself*. But we submit that in the limited information world of practical policy making, such uncertainties are rife.[2]

We illustrate the concept of a policy lottery using the CGE model documented in Harrison et al. (2002). This static model of the Danish economy is calibrated to data from 1992. The version we use has 27 production sectors, each employing intermediate inputs and primary factors to produce output for domestic and overseas consumption. A government agent raises taxes and pays subsidies in a revenue-neutral manner, and the focus of our policy simulation is on the indirect taxes levied by the Danish government.[3] A representative government household consumes goods reflecting public expenditure patterns in 1992. The simulated policy effects are different across several private household types. The model is calibrated to a wide array of empirical and a priori estimates of elasticities of substitution using nested constant elasticity of substitution specifications for production and utility functions. More elaborate versions of the model exist in which intertemporal and intergenerational behavior are modeled (e.g., Lau, 2000), but this static version is ideal for our illustrative purposes.

The model represents several different private households, based on the breakdown provided by Statistics Denmark from the national household expenditure survey. For our purposes, these households are differentiated by family type into seven households: singles younger than 45 without children, singles older than 45 without children, households younger than 45 without children, households older than 45 without children, singles with children, households with children and where the oldest child is 6 or under, and households with children and where the oldest child is between 7 and 17. The model generates the welfare impact on each of these households measured in terms of the equivalent variation in annual income for that household. That is, it calculates the amount of income the household would deem to be equivalent to the policy change, which entails changes in factor prices, commodity prices, and expenditure patterns. Thus the policy impact is some number of Danish kroner, which represents the welfare gain to the household in income terms.

This welfare gain can be viewed directly as the "prize" in a policy lottery. Since there is some uncertainty about the many parameters used to calibrate realistic simulation models of this kind, there is some uncertainty about the calculation of the welfare impact. If we perturb one or more of the elasticities, for example, the

welfare gain might well be above or below the baseline computation. Using randomized factorial designs for such sensitivity analyses, we can undertake a large number of these perturbations and assign a probability weight to each one (Harrison and Vinod, 1992). Each simulation involves a random draw for each elasticity, but where the value drawn reflects estimates of the empirical distribution of the elasticity.[4] We undertake 1000 simulations with randomly generated elasticity perturbations, so it is as if the household faces a policy lottery consisting of 1000 distinct prizes that occur with equal probability 0.001. The prizes, again, are the welfare gains that the model solves for in each such simulation.

Figure 15.1 illustrates the type of policy lottery that can arise. In this case we consider a policy of making all indirect taxes in Denmark uniform, and at a uniform value that just maintains the real value of government expenditure. Thus we solve for a revenue-neutral reform in which the indirect tax distortions arising from inter-sectoral variation in those taxes are reduced to zero. Each box in Figure 15.1 represents 1000 welfare evaluations of the model for each household type. The large dot is the median welfare impact, the rectangle is the interquartile range,[5] and the whiskers represent the range of observed values. Thus we see that the policy represents a lottery for each household, with some uncertainty about the impacts.

Generation of policy lotteries are not restricted to CGE models. The method applies to any simulation model that generates outcomes that reflect policy changes. For example, Fiore et al. (2009) used a simulation model of the spread of forest fire, developed by the USDA for that purpose and calibrated to detailed

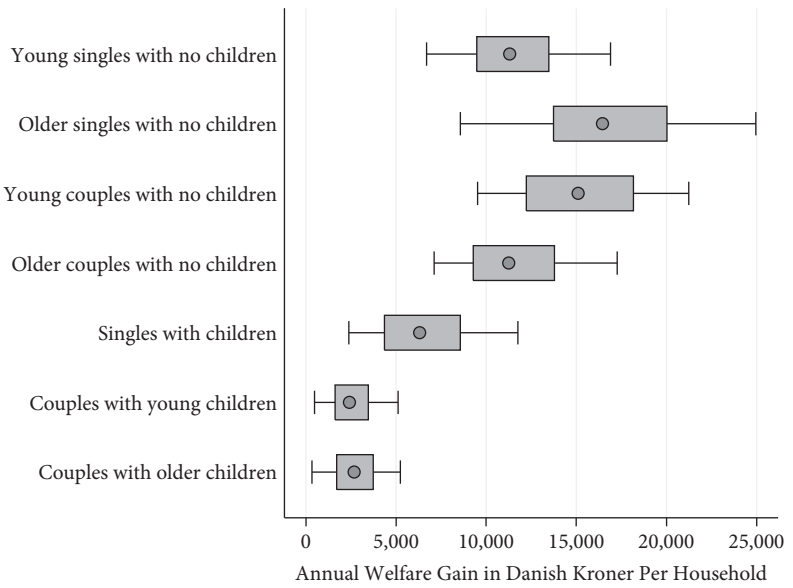


Figure 15.1. An illustrative policy lottery. Distribution of welfare effects of indirect tax uniformity.

GIS data for a specific area, to generate policy lotteries for experimental subjects to make choices over. Our approach just recognizes that policy models of this kind are never certain and that they contain standard errors—in fact, lots of standard errors. But that uncertainty should not be ignored when the policy maker uses the model to decide on good policies.

The idea that policies are lotteries is a simple one, and it is well known in the older simulation literature in CGE modeling. The methods developed to address it amounted to Monte Carlo analyses on repeated simulations in which each uncertain parameter was perturbed around its point estimate. By constraining these perturbations to within some empirical or a priori confidence region, one implicitly constrained the simulated policy outcome to that region. The same idea plays a central role in the *Stern Review on the Economics of Climate Change* (Stern, 2007). It stresses (p. 163) the need to have a simulation model of the economic effects of climate change that can show stochastic impacts. In fact, any of the standard climate simulation models can easily be set up to do that, by simply undertaking a systematic sensitivity analysis of their results. The *Review* then proposes an "expected utility analysis" of the costs of climate change (p. 173ff), which is effectively the same as viewing climate change impacts as a lottery. When one then considers alternative policies to mitigate the risk of climate change, the "expected utility analysis" is the same as our policy lottery concept.

If a policy maker were to evaluate the expected utility to each household from this policy, he would have to take into account the uncertainty of the estimated outcome and the risk attitudes of the household. The traditional approach in policy analysis is to implicitly assume that households are all risk neutral and simply report the average welfare impact. But we know from our experimental results that these households are not risk neutral. Assume a Constant Relative Risk Aversion (CRRA) utility specification for each household. Anticipating the later discussion of our experimental results, we can stratify our raw elicited CRRA intervals according to these seven households and obtain CRRA estimates of 1.17, 0.48, 0.79, 0.69, 0.76, 0.81, and 0.95, respectively, for each of these households. In each case these are statistically significantly different from risk neutrality.

Using these CRRA risk attitude estimates, it is a simple matter to evaluate the utility of the welfare gain in each simulation, to then calculate the expected utility of the proposed policy, and to finally calculate the certainty-equivalent welfare gain. Doing so reduces the welfare gain relative to the risk-neutral case, of course, since there is some uncertainty about the impacts. For this illustrative policy, this model, these empirical distributions of elasticities, and these estimates of risk attitudes, we find that the neglect of risk aversion results in an overstatement of the welfare gains by 1.6%, 1.4%, 1.8%, 1.1%, 5.1%, 4.6%, and 7.9%, respectively, for each of the households. Thus a policy maker would overstate the welfare gains from the policy if risk attitudes were ignored.

Tax uniformity is a useful pedagogic example, as well as a staple in public economics, but one that generates relatively precise estimates of welfare gains in most

simulation models of this kind. It is easy to consider alternative realistic policy simulations that would generate much more variation in welfare gain, and hence larger corrections from using the household's risk attitude in policy evaluation. For example, assume instead that indirect taxes in this model were reduced across the board by 25%, and that the government affected lump-sum side payments to each household to ensure that no household had less than a 1% welfare gain.[6] In this case, plausible elasticity configurations for the model exist that result in very large welfare gains for some households.[7] Ignoring the risk attitudes of the households would result in welfare gains being overstated by much more significant amounts, ranging from 18.9% to 42.7% depending on the household.

These policy applications point to the payoff from estimating risk attitudes, as we do here, but they are only illustrative. A number of limiting assumptions obviously have to be imposed on our estimates for them to apply to the policy exercise. First, we have to assume that the estimates of CRRA obtained from our experimental tasks defined over the domain of prizes up to 4500 DKK apply more widely, to the domain of welfare gains shown in Figure 15.1. Given the evidence from our estimation of the Expo-Power function, reported in Harrison et al. (2007), we are prepared to make that assumption for now. Obviously one would want to elicit risk attitudes over wider prize domains to be confident of this assumption, however. Second, we only aggregate households into seven different types, each of which is likely to contain households with widely varying characteristics on other dimensions than family types. Despite these limitations, these illustrations point out the importance of attending to the risk preference assumptions imposed in policy evaluations. Recent efforts in modeling multiple households in computable general equilibrium have been driven by concerns about the impacts of trade reform on poverty in developing countries, since one can only examine those by identifying the poorest households (see Harrison et al. (2003) and Harrison et al. (2004)). Clearly one would expect risk aversion to be a particularly important factor for households close to or below the absolute poverty line.

It might be apparent that we would have to conduct field experiments with a sample representative of the Danish population in order to calibrate a CGE model of the Danish economy to risk attitudes that were to be regarded as having any credibility with policy makers. But perhaps this is not so obvious to academics, who are often happy to generalize from convenience samples. In a related setting, in this instance with respect to behavioral findings from laboratory experiments that question some of the theoretical foundations of welfare economics, List (2005, p. 36) records that in his

> . . . discussions with agency officials in the U.S. who perform/oversee benefit–cost analyses, many are aware of these empirical findings, and realize that they have been robust across unfamiliar goods, such as irradiated sandwiches, and common goods, such as chocolate bars, but many remain skeptical of the received results. Most importantly for our purposes, some policymakers view experimental laboratory results with a degree of suspicion, one noting that the methods are

> akin to "scientific numerology." When pressed on this issue, some suggest that their previous experience with stated preference surveys leads them to discount experimental results, especially those with student samples, and they conclude that the empirical findings do not merit policy changes yet. A few policy officials openly wondered if the anomalous findings would occur in experiments with "real" people.

Our experience has been the same and is the reason why we were led to conduct field experiments in Denmark.

Risk Aversion

In order to evaluate the policy lottery considered in the previous section, we needed to have estimates of the risk attitudes for the different households in Denmark. We therefore designed an experiment to elicit risk attitudes (and discount rates) from representative Danes. The experiment is a longitudinal panel where we revisited many of the first-stage participants at a later date. In this section we discuss the issues that arose in our field experiments, with an emphasis on those issues that are relatively novel as a result of the field context.

The immediate implication, of course, was that we needed to generate a sampling frame that allowed us to make inferences about the broader adult population in Denmark. This led us to employ stratified sampling methods for large-scale surveys, which are relatively familiar to labor economists and health economists but which had not been used in the experimental literature. We were also concerned about possible sample selection effects from our recruiting strategy, as well as the possibility of what is known in the literature as "randomization bias." Two types of sample selection effects were possible. First, we were concerned that the information about earnings in the recruitment information would attract a sample biased in the direction of risk loving. Second, we were concerned that particular experiences in the first stage of the experiment could bias attrition to the second stage. These concerns not only influenced our econometric strategy but also led to the design of complementary lab experiments to directly test for such effects.

The next concern was with the design of the elicitation procedure itself. There were many alternatives available in the literature, along with known trade-offs from using one or the other. We were particularly concerned to have an elicitation procedure that could be relatively easily implemented in the field, even though we had the benefit compared to some field contexts of being able to assume a literate population. We use elicitation procedures that do not have a specific context since the purpose was to generate risk preference parameters for general policy use. We use complementary lab experiments to condition our field inferences on any vulnerability in responses to variations in procedures. These procedural variations were guided by hypotheses about the effect of frames on the participants' perception of the task and on their use of information processing heuristics.

Once we had collected the experimental data, several issues arose concerning the manner in which one infers the risk attitudes. These issues demanded the use of an explicit, structural approach to estimating models of choice over risky lotteries. The reason is that we wanted to obtain estimates of the latent parameters of these choice models and to be able to evaluate alternative choice models at a structural level. One attraction of this approach is that it allowed us to be explicit about issues that are often left implicit but which can have a dramatic affect on inferred risk attitudes; one example is the specification of what is known as a "behavioral error term" in these choice models. Another attraction is that it allowed us to examine alternative theories to expected utility theory using a comparable inferential framework.

Our goal was to generate measures of risk attitudes for a range of monetary prizes and over time. With this data we can investigate the robustness of the measures over time, as reflective of stationary or state dependent preferences, and robustness with respect to income changes.

Sampling Procedures

The sample for the field experiments was designed to be representative of the adult Danish population in 2003. There were six steps in the construction of the sample, detailed in Harrison et al. (2005) and essentially following those employed in Harrison et al. (2002):

- First, a random sample of 25,000 Danes was drawn from the Danish Civil Registration Office in January 2003. Only Danes born between 1927 and 1983 were included, thereby restricting the age range of the target population to between 19 and 75. For each person in this random sample we had access to their name, address, county, municipality, birth date, and sex. Due to the absence of names and/or addresses, 28 of these records were discarded.
- Second, we discarded 17 municipalities (including one county) from the population, due to them being located in extraordinarily remote locations and hence being very costly to recruit. The population represented in these locations amounts to less than 2% of the Danish population, or 493 individuals in our sample of 25,000 from the Civil Registry. Hence it is unlikely that this exclusion could quantitatively influence our results on sample selection bias.
- Third, we assigned each county either 1 session or 2 sessions, in rough proportionality to the population of the county. In total we assigned 20 sessions. Each session consisted of two sub-sessions at the same locale and date, one at 5 pm and another at 8 pm, and subjects were allowed to choose which sub-session suited them best.
- Fourth, we divided six counties into two subgroups because the distance between some municipalities in the county and the location of the session

would be too large. A weighted random draw was made between the two subgroups and the location selected, where the weights reflect the relative size of the population in September 2002.

- Fifth, we picked the first 30 or 60 randomly sorted records within each county, depending on the number of sessions allocated to that county. This provided a subsample of 600.
- Sixth, we mailed invitations to attend a session to the subsample of 600, offering each person a choice of times for the session. Response rates were low in some counties, so another 64 invitations were mailed out in these counties to newly drawn subjects.[8] Everyone that gave a positive response was assigned to a session, and our recruited sample was 268.

Attendance at the experimental sessions was extraordinarily high, including four persons who did not respond to the letter of invitation but showed up unexpectedly and participated in the experiment. Four persons turned up for their session, but were not able to participate in the experiments.[9] These experiments were conducted in June of 2003, and a total of 253 subjects participated.[10] Sample weights for the subjects in the experiment can be constructed using this experimental design, and they can be used to calculate weighted distributions and averages that better reflect the adult population of Denmark.

Elicitation Procedures

There are many general elicitation procedures that have been used in the literature to ascertain risk attitudes from individuals in the experimental laboratory using noninteractive settings, and each is reviewed in detail by Harrison and Rutström (2008). Most of these simply present participants with lotteries specified using various monetary prizes and probabilities without attaching a particular context; these are labeled "artifactual" presentations. An approach made popular by Holt and Laury (2002) is the Multiple Price List (MPL), which entails giving the subject an ordered array of binary lottery choices to make all at once. The MPL requires the subject to pick one of the lotteries being offered, and then it plays that lottery out for the subject to be rewarded. The earliest use of the MPL design in the context of elicitation of risk attitudes is, we believe, Miller et al. (1969). Their design confronted each subject with five alternatives that constitute an MPL, although the alternatives were presented individually over 100 trials. The method was later used by Schubert et al. (1999), Barr and Packard (2002), and, of course, Holt and Laury (2002). The MPL has the advantage of allowing the subject to easily compare options involving various risks. As is the case with all procedures of this nature, there is some question about the robustness of responses with respect to procedural variations. We decided to use complementary lab experiments to explore several of these procedural issues, rather than incur the expense of evaluating them in the field.

In our field version of the MPL, each subject is presented with a choice between two lotteries, which we can call A or B. Table 15.1 illustrates the basic payoff matrix presented to subjects in our experiments. The complete procedures are described in Harrison et al. (2005). The first row shows that lottery A offered a 10% chance of receiving 2000 DKK and a 90% chance of receiving 1600 DKK. The expected value of this lottery, EV^A, is shown in the third-last column as 1640 DKK, although the EV columns were not presented to subjects. Similarly, lottery B in the first row has chances of payoffs of 3850 and 100 DKK, for an expected value of 475 DKK. Thus the two lotteries have a relatively large difference in expected values, in this case 1165 DKK. As one proceeds down the matrix, the expected value of both lotteries increases, but the expected value of lottery B becomes greater relative to the expected value of lottery A.

In a traditional MPL the subject chooses A or B in each row, and one row is later selected at random for payout for that subject. The logic behind this test for risk aversion is that only risk-loving subjects would take lottery B in the first row, and only very risk-averse subjects would take lottery A in the second last row.[11] Arguably, the last row is simply a test where the subject understood the instructions, and it has no relevance for risk aversion at all. A risk-neutral subject should switch from choosing A to B when the EV of each is about the same, so a risk-neutral subject would choose A for the first four rows and B thereafter. In our field implementation we instead had the subject choose on which row to switch from A to B, thus forcing monotonicity, but we also added an option to indicate indifference; we refer to this variant of the MPL as a sequential MPL (sMPL). For those subjects who did not express indifference, we recognized the opportunity to get more refined measures by following up with a subsequent stage where the probabilities attached to the prizes lay within the range of those on the previous switching interval; we refer to this variant as the Iterative MPL (iMPL).[12]

The iMPL uses the same incentive logic as the MPL and sMPL. The logic of selecting a row for payment is maintained but necessitated a revision of the random method used. Let the first stage of the iMPL be called Level 1, the second stage Level 2, and so on. After making all responses, the subject has one row from the first table of responses in Level 1 selected at random by the experimenter. In the MPL, that is all there is since there is only a Level 1 table. In the iMPL, that is all there is if the row selected at random by the experimenter is *not* the one at which the subject switched in Level 1. If it *is* the row at which the subject switched, another random draw is made to pick a row in the Level 2 table. For some tasks this procedure is repeated to Level 3.

In order to investigate what effect there may be on responses from using the iMPL, we ran lab experiments comparing this procedure to the standard MPL and the sMPL (see Andersen et al. (2006)). As noted above, the sMPL changes the MPL to ask the subject to pick the switch point from one lottery to the other, but without the refinement of probabilities allowed in iMPL. Thus it enforces monotonicity, but still allows subjects to express indifference at the "switch" point, akin to a "fat switch

Table 15.1. Typical Payoff Matrix in the Danish Risk Aversion Experiments

Lottery A				Lottery B				EV^A	EV^B	Difference	Open CRRA Interval if Subject Switches to Lottery B and $\omega = 0$
p	DKK	p	DKK	p	DKK	p	DKK	DKK	DKK	DKK	
0.1	2000	0.9	1600	0.1	3850	0.9	100	1640	475	1165	$-\infty$, –1.71
0.2	2000	0.8	1600	0.2	3850	0.8	100	1680	850	830	–1.71, –0.95
0.3	2000	0.7	1600	0.3	3850	0.7	100	1720	1225	495	–0.95, –0.49
0.4	2000	0.6	1600	0.4	3850	0.6	100	1760	1600	160	–0.49, –0.15
0.5	2000	0.5	1600	0.5	3850	0.5	100	1800	1975	–175	–0.15, 0.14
0.6	2000	0.4	1600	0.6	3850	0.4	100	1840	2350	–510	0.14, 0.41
0.7	2000	0.3	1600	0.7	3850	0.3	100	1880	2725	–845	0.41, 0.68
0.8	2000	0.2	1600	0.8	3850	0.2	100	1920	3100	–1180	0.68, 0.97
0.9	2000	0.1	1600	0.9	3850	0.1	100	1960	3475	–1515	0.97, 1.37
1	2000	0	1600	1	3850	0	100	2000	3850	–1850	1.37, ∞

Note: The last four columns in this table, showing the expected values of the lotteries and the implied CRRA intervals, were not shown to subjects.

point." The subject was then paid in the same manner as with MPL, but with the non-switch choices filled in automatically.

We used four separate risk aversion tasks with each subject, each with different prizes designed so that all 16 prizes span the range of income over which we seek to estimate risk aversion. The four sets of prizes were as follows, with the two prizes for lottery A listed first and the two prizes for lottery B listed next: (A1: 2000 DKK, 1600 DKK; B1: 3850 DKK, 100 DKK), (A2: 2250 DKK, 1500 DKK; B2: 4000 DKK, 500 DKK), (A3: 2000 DKK, 1750 DKK; B3: 4000 DKK, 150 DKK), and (A4: 2500 DKK, 1000 DKK; B4: 4500 DKK, 50 DKK). At the time of the experiments, the exchange rate was approximately 6.55 DKK per U.S. dollar, so these prizes ranged from approximately $7.65 to $687.

We ask the subject to respond to all four risk aversion tasks and then randomly decide which task and row to play out. In addition, the large incentives and budget constraints precluded paying all subjects, so each subject is given a 10% chance to actually receive the payment associated with his decision.

We take each of the binary choices of the subject as the data, and we estimate the parameters of a latent utility function that explains those choices using an appropriate error structure to account for the panel nature of the data. Once the utility function is defined, for a candidate value of the parameters of that function, we can construct the expected utility of the two gambles and then use a linking function to infer the likelihood of the observed choice. We discuss statistical specifications in more detail below.

The MPL instrument has an apparent weakness because it might suggest a frame that encourages subjects to select the middle row, contrary to their unframed risk preferences. The antidote for this potential problem is to devise various "skewed" frames in which the middle row implies different risk attitudes, and then see if there are differences across frames. Simple procedures to detect such framing effects, and correct for them statistically if present, have been developed and are discussed below (e.g., Harrison et al., 2005; Andersen et al., 2006; and Harrison et al., 2007).

In summary, the set of MPL instruments provides a relatively transparent procedure to elicit risk attitudes. Subjects rarely get confused about the incentives to respond truthfully, particularly when the randomizing devices are physical die that they know that they will toss themselves.[13] As we demonstrate later, it is also possible to infer a risk attitude interval for the specific subject, at least under some reasonable assumptions, as well as to use the choice data to estimate structural parameters of choice models.

Estimation Procedures

Two broad methods of estimating risk attitudes have been used. One involves the calculation of bounds implied by observed choices, typically using utility functions which only have a single parameter to be inferred. A major limitation of this

approach is that it restricts the analyst to utility functions that can characterize risk attitudes using one parameter. This is because one must infer the bounds that make the subject indifferent between the switch points, and such inferences become virtually incoherent statistically when there are two or more parameters. Of course, for popular functions such as CRRA or Constant Absolute Risk Aversion (CARA), this is not an issue; but if one wants to move beyond those functions, then there are problems. It is possible to devise one-parameter functional forms with more flexibility than CRRA or CARA in some dimension, as illustrated nicely by the one-parameter Expo-Power function developed by Abdellaoui et al. (2007, Section 4). But in general we need to move to structural modeling with maximum likelihood to accommodate richer models.

The other broad approach involves the direct estimation by maximum likelihood of some structural model of a latent choice process in which the core parameters defining risk attitudes can be estimated, in the manner pioneered by Camerer and Ho (1994, Section 6.1) and Hey and Orme (1994). This structural approach is particularly attractive for non-EUT specifications, where several core parameters combine to characterize risk attitudes. For example, one cannot characterize risk attitudes under Prospect Theory without making some statement about loss aversion and probability weighting, along with the curvature of the utility function. Thus joint estimation of all parameters is a necessity for reliable statements about risk attitudes in such cases.[14]

Assume for the moment that utility of income is defined by

$$U(x) = \frac{x^{(1-r)}}{1-r}, \tag{15.1}$$

where x is the lottery prize and $r \neq 1$ is a parameter to be estimated. For $r = 1$ assume $U(x) = \ln(x)$ if needed. We come back later to the controversial issue of what x might be, but for now we assume that it is just the monetary prize M shown in the lottery. Thus r is the coefficient of CRRA: $r = 0$ corresponds to risk neutrality, $r < 0$ corresponds to risk loving, and $r > 0$ corresponds to risk aversion. Let there be two possible outcomes in a lottery. Under Expected Utility Theory (EUT) the probabilities for each outcome M_j, $p\left(M_j\right)$, are those that are induced by the experimenter, so expected utility is simply the probability weighted utility of each outcome j in each lottery i plus some level of background consumption ω:

$$EU_i = \sum_{j=1,2} \left[p\left(M_j\right) \times U\left(\omega + M_j\right) \right]. \tag{15.2}$$

The EU for each lottery pair is calculated for a candidate estimate of r, and the index

$$\nabla EU = EU_R - EU_L \tag{15.3}$$

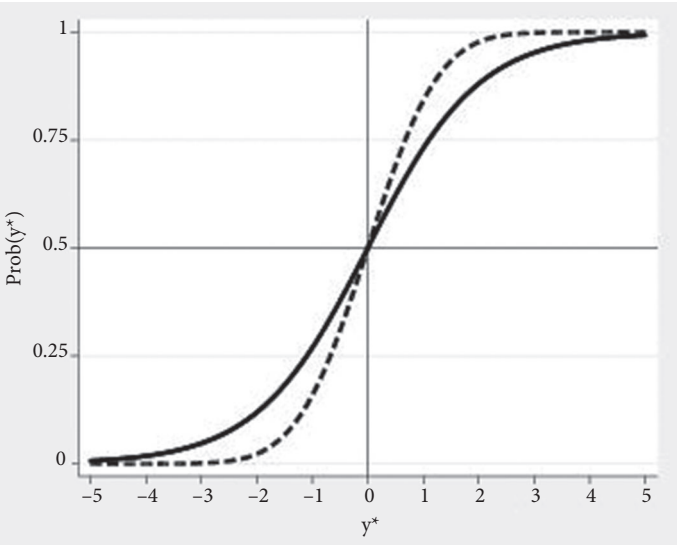


Figure 15.2. Normal and logistic cumulative density functions. Dashed line is normal, and solid line is logistic.

is calculated, where EU_L is the "left" lottery and EU_R is the "right" lottery as presented to subjects. This latent index, based on latent preferences, is then linked to observed choices using a standard cumulative normal distribution function $\Phi(\nabla \text{EU})$. This "probit" function takes any argument between $\pm\infty$ and transforms it into a number between 0 and 1 using the function shown in Figure 15.2. Thus we have the probit link function,

$$\text{prob(choose lottery } R) = \Phi(\nabla EU). \tag{15.4}$$

The logistic function is very similar, as illustrated in Figure 15.2, and leads instead to the "logit" specification.

Even though Figure 15.2 is common in econometrics texts, it is worth noting explicitly and understanding. It forms the critical statistical link between observed binary choices, the latent structure generating the index $y^* = \nabla EU$, and the probability of that index y^* being observed. In our applications, y^* refers to some function, such as (15.3), of the EU of two lotteries; or, if one is estimating a Prospect Theory model, the prospective utility of two lotteries. The index defined by (15.3) is linked to the observed choices by specifying that the R lottery is chosen when $\Phi(\nabla EU) > \frac{1}{2}$, which is implied by (15.4).

Thus the likelihood of the observed responses, conditional on the EUT and CRRA specifications being true, depends on the estimates of r given the above statistical specification and the observed choices. The "statistical specification" here includes assuming some functional form for the cumulative density function

(CDF), such as one of the two shown in Figure 15.2. If we ignore responses that reflect indifference for the moment, the conditional log-likelihood would be

$$\ln L(r; y, \omega, \mathbf{X}) = \sum_i \left[\left(\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1) \right) + \left(\ln(1 - \Phi(\nabla EU)) \times \mathbf{I}(y_i = -1) \right) \right], \tag{15.5}$$

where $\mathbf{I}(\cdot)$ is the indicator function, $y_i = 1(-1)$ denotes the choice of the option $R(L)$ lottery in risk aversion task i, and $\mathbf{X}$ is a vector of individual characteristics reflecting age, sex, race, and so on. The parameter r is defined as a linear function of the characteristics in vector $\mathbf{X}$.

In most experiments the subjects are told at the outset that any expression of indifference would mean that if that choice was selected to be played out, a fair coin would be tossed to make the decision for them. Hence one can modify the likelihood to take these responses into account by recognizing that such choices implied a 50:50 mixture of the likelihood of choosing either lottery:

$$\begin{aligned} \ln L(r; y, \omega, \mathbf{X}) = \sum_i \Bigg[& \left(\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1) \right) + \left(\ln(1 - \Phi(\nabla EU)) \times \mathbf{I}(y_i = -1) \right) \\ & + \left(\left(\frac{1}{2} \ln \Phi(\nabla EU) + \frac{1}{2} \ln (1 - \Phi(\nabla EU)) \right) \times \mathbf{I}(y_i = 0) \right) \Bigg], \end{aligned} \tag{15.5′}$$

where $y_i = 0$ denotes the choice of indifference. In our experience, very few subjects choose the indifference option, but this formal statistical extension accommodates those responses.[15]

The latent index (15.3) could have been written in a ratio form:

$$\nabla EU = \frac{EU_R}{EU_R - EU_L}, \tag{15.3′}$$

and then the latent index would already be in the form of a probability between 0 and 1, so we would not need to take the probit or logit transformation. This specification has also been used, with some modifications we discuss later, in Holt and Laury (2002).

Harrison and Rutström (2008, Appendix F) review procedures and syntax from the popular statistical package *Stata* that can be used to estimate structural models of this kind, as well as more complex non-EUT models. The goal is to illustrate how experimental economists can write explicit maximum likelihood (ML) routines that are specific to different structural choice models. It is a simple matter to correct for stratified survey responses, multiple responses from the same subject ("clustering"),[16] or heteroskedasticity, as needed.

Using the CRRA utility function and equations (15.1) through (15.4), we estimate r to be 0.78 for the Danish population, with a standard error of 0.052 and a 95% confidence interval between 0.68 and 0.88. This reflects modest risk aversion over these stakes and is significantly different from risk neutrality ($r = 0$).

Extensions of the basic model are easy to implement, and this is the major attraction of the structural estimation approach. For example, one can easily extend the functional forms of utility to allow for varying degrees of relative risk aversion (RRA). Consider, as one important example, the Expo-Power (EP) utility function proposed by Saha (1993). Following Holt and Laury (2002), the EP function is defined as

$$U(x) = \frac{1 - \exp(-\alpha x^{1-r})}{\alpha}, \tag{15.1′}$$

where α and r are parameters to be estimated. RRA is then $r + \alpha(1 - r)y^{1-r}$, so RRA varies with income if $\alpha \neq 0$. This function nests CRRA (as $\alpha \rightarrow 0$) and CARA (as $r \rightarrow 0$). We illustrate the use of this EP specification in Harrison et al. (2007).

It is also a simple matter to generalize this ML analysis to allow the core parameter r to be a linear function of observable characteristics of the individual or task. We would then extend the model to be $r = r_0 + \mathbf{R} \times \mathbf{X}$, where r_0 is a fixed parameter and $\mathbf{R}$ is a vector of effects associated with each characteristic in the variable vector $\mathbf{X}$. In effect the unconditional model assumes $r = r_0$ and just estimates r_0. This extension significantly enhances the attraction of structural ML estimation, particularly for responses pooled over different subjects, since one can condition estimates on observable characteristics of the task or subject.

An important extension of the core model is to allow for subjects to make some errors. The notion of error is one that has already been encountered in the form of the statistical assumption that the probability of choosing a lottery is not 1 when the *EU* of that lottery exceeds the *EU* of the other lottery. This assumption is clear in the use of a link function between the latent index ∇EU and the probability of picking one or other lottery; in the case of the normal CDF, this link function is $\Phi(\nabla EU)$ and is displayed in Figure 15.2. If there were no errors from the perspective of EUT, this function would be a step function, which is shown in Figure 15.3: zero for all values of $y^* < 0$, anywhere between 0 and 1 for $y^* = 0$, and 1 for all values of $y^* > 0$.

The problem with the CDF of the Hardnose Theorist is immediate: It predicts with probability one or zero. The likelihood approach asks the model to state the probability of observing the actual choice, conditional on some trial values of the parameters of the theory. Maximum likelihood then locates those parameters that generate the highest probability of observing the data. For binary choice tasks and independent observations, we know that the likelihood of the sample is just the product of the likelihood of each choice conditional on the model and the parameters assumed and that the likelihood of each choice is just the probability of that

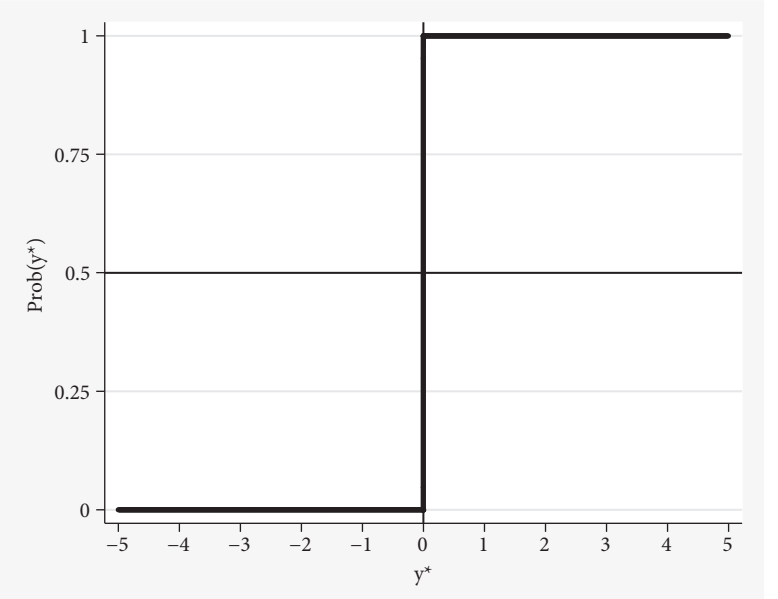


Figure 15.3. Hardnose theorist cumulative density function.

choice. So if we have any choice that has zero probability, and it might be literally one-in-a-million choices, the likelihood for that observation is not defined. Even if we set the probability of the choice to some arbitrarily small, positive value, the log-likelihood zooms off to minus infinity. We can reject the theory without even firing up any statistical package.

Of course, this implication is true for any theory that predicts deterministically, including Expected Utility Theory (EUT). This is why one needs some formal statement about how the deterministic prediction of the theory translates into a probability of observing one choice or the other, and then perhaps also some formal statement about the role that structural errors might play.[17] In short, one *cannot divorce the job of the theorist from the job of the econometrician*, and some assumption about the process linking latent preferences and observed choices is needed. That assumption might be about the mathematical form of the link, as in (15.1), but it cannot be avoided. Even the very definition of risk aversion needs to be specified using stochastic terms unless we are to impose absurd economic properties on estimates (Wilcox, 2008, 2010).

By varying the shape of the link function in Figure 15.2, one can informally imagine subjects that are more sensitive to a given difference in the index ∇EU and subjects that are not so sensitive. Of course, such informal intuition is not strictly valid, since we can choose any scaling of utility for a given subject, but it is suggestive of the motivation for allowing for structural errors and why we might want them to vary across subjects or task domains.

Consider the structural error specification used by Holt and Laury (2002), originally due to Luce. The *EU* for each lottery pair is calculated for candidate estimates of *r*, as explained above, and the ratio

$$\nabla EU = \frac{EU_R^{1/\mu}}{EU_L^{1/\mu} + EU_R^{1/\mu}} \tag{15.3''}$$

calculated, where μ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model. The index ∇EU is in the form of a cumulative probability distribution function defined over the noise parameter μ and the ratio of the *EU* of one lottery to the sum of the *EU* of both lotteries. Thus, as $\mu \rightarrow 0$ this specification collapses to the deterministic choice EUT model, where the choice is strictly determined by the *EU* of the two lotteries; but as μ gets larger and larger, the choice essentially becomes random. When $\mu = 1$, this specification collapses to (15.3′), where the probability of picking one lottery is given directly by the ratio of the *EU* of one lottery to the sum of the *EU* of both lotteries. Thus μ can be viewed as a parameter that flattens out the link functions in Figure 15.2 as it gets larger. This is just one of several different types of error story that could be used, and Wilcox (2008) provides a masterful review of the implications of the alternatives.[18]

There is one other important error specification, due originally to Fechner and popularized by Hey and Orme (1994). This error specification posits the latent index

$$\nabla EU = \frac{EU_R - EU_L}{\mu} \tag{15.3'''}$$

instead of (15.3), (15.3′) or (15.3″).

Wilcox (2008) notes that as an analytical matter the evidence of IRRA in Holt and Laury (2002) would be weaker, or perhaps even absent, if one had used a Fechner error specification instead of a Luce error specification. This important claim, that the evidence for IRRA may be an artifact of the (more or less arbitrary) stochastic identifying restriction assumed, can be tested with the original data from Holt and Laury (2002) and is correct (see Harrison and Rutström (2008, Figure 9)).

An important contribution to the characterization of behavioral errors is the "contextual error" specification proposed by Wilcox (2010). It is designed to allow robust inferences about the primitive "more stochastically risk averse than" and avoids the type of "residual-tail-wagging-the-dog" results that one gets when using the Fechner or Luce specification and the Holt and Laury (2002) data. It posits the latent index

$$\nabla EU = \frac{(EU_R - EU_L)\nu}{\mu} \tag{15.3*}$$

instead of (15.3′′′), or

$$\nabla EU = \frac{(EU_R/\nu)^{1/\mu}}{(EU_L/\nu)^{1/\mu} + (EU_R/\nu)^{1/\mu}} \tag{15.3**}$$

instead of (15.3′′), where ν is a new, normalizing term for each lottery pair L and R. The normalizing term ν is defined as the maximum utility over all prizes in this lottery pair minus the minimum utility over all prizes in this lottery pair. The value of ν varies, in principle, from lottery choice; to lottery choice; hence it is said to be "contextual." For the Fechner specification, dividing by ν ensures that the *normalized EU* difference $[(EU_R - EU_L)/\nu]$ remains in the unit interval.

Asset Integration

There is a tension between experimental economists and theorists over the proper interpretation of estimates of risk attitudes that emerge from experimental choice behavior. Experimental economists claim to provide evidence of risk aversion over small stakes, which we take here to be amounts such as $10, $100, or even several hundred dollars. Some theorists argue that these estimates are "implausible," in a sense to be made explicit. Although the original arguments of theorists were couched as attacks on the plausibility of EUT (e.g., Hansson, 1988; Rabin, 2000), it is now apparent that the issues are just as important, or unimportant, for non-EUT models (e.g., Safra and Segal, 2008; Cox and Sadiraj, 2008).

The notion of plausibility can be understood best by thinking of the argument in several steps, even if they are often collapsed into one. First, someone proposes *point* estimates of some utility function. These estimates typically come from inferences based on observed choice behavior in an experiment, derived from maximum likelihood estimates of a structural model of latent choice behavior (e.g., Harrison and Rutström, 2008, Section 3). Standard errors on those estimates are usually not relied on in these exercises. Second, with some auxiliary assumptions, an analyst constructs a lottery choice task in which these point estimates generate predictions on some domain of lottery prizes, typically involving at least one prize that is much larger than the domain of prizes over which the estimates were derived (e.g., Rabin, 2000; Cox and Sadiraj, 2008, Section 3.2). In some cases these predictions are also defined on the domain of lottery prizes of the experimental tasks (e.g., Cox and Sadiraj, 2008, Section 4.5). Third, the analyst views this constructed lottery choice task as a "thought experiment," in the sense that it is just like an actual experiment except that it is not actually implemented (Harrison and List, 2004, Section 9). One reason not to conduct the experiment is that it might involve astronomic stakes, but the main reason is that it is assumed a priori obvious what the choice would be, in some sense eliminating the need for additional experiments. Finally, it is pointed out that the predicted outcome from the initial estimates is contrary to the a priori obvious choice in the thought experiment. Thumbs down, and the initial estimates are discarded as implausible.

Since the initial estimates are typically defined over observed experimental choices with small stakes, we refer here to the implied claims about risk aversion as "risk aversion in the small." The predicted behavior in the thought experiment is typically defined over choices with very large stakes, so we refer to the implied choices as reflecting "risk aversion in the large." So plausibility can be viewed as a tension and inconsistency between observations (real and imagined a priori) generated on two domains. The issue is not that the subject has to have the same relative or absolute measure of risk aversion for different prizes: these problems arise even when "flexible" functional forms are employed for utility functions.

One general response is to just focus on risk attitudes in the small and to make no claims about behavior beyond the domain over which the estimates were obtained. This position states that if one had estimated over larger domains, then the estimated models would reflect actual choices over that domain, but one simply cannot apply the risk aversion estimates outside the domain of estimation. Since large parts of economic theory are written in terms of the utility of income, rather than the utility of wealth, this approach has some validity. Of course, there is nothing in principle to stop one from defining income as a large number, either with a large budget, with subjects in a very poor country (e.g., Harrison et al., 2010), or by using natural experiments such as game shows (e.g., Andersen et al., 2008c).

A second approach, which we employed in Andersen et al. (2008a), was to assume some level of baseline consumption that was suggested by expenditure data for the subjects.

A third approach is to test for the degree of asset integration in observed behavior. If one adopts a general specification, following Cox and Sadiraj (2006), and allows income and wealth to be arguments of some utility function, then one does not have to assume that the argument of the utility function is income or wealth. One might posit an aggregation function that combines the two in some way, with this composite then being evaluated with some standard utility function. For example, assume the linear aggregation function $\omega W + y$, where W is wealth, y is experimental income, and ω is some weighting parameter to be assumed or estimated. Or one could treat ωW and y as inputs into some Constant Elasticity of Substitution function, as well as estimate or assume ω and the elasticity of substitution. This approach allows the popular special cases of zero asset integration and perfect asset integration, but lets the "data decide" when these parameters are estimated in the presence of actual choices. Where does one get estimates of W? As it happens, very good proxies for W can be inferred from data in Denmark that is collected by *Statistics Denmark*, the official government statistics agency. One can then calculate those proxies for subjects that have been in experiments such as ours (such things being feasible in Denmark, with appropriate confidentiality agreements) and estimate the weighting parameter. The evidence from Andersen et al. (2011) suggests that ω is very small indeed and that the elasticity of substitution between ωW and y is close to 1.[19]

Discount Rates

In many settings in experimental economics, we want to elicit some preference from a set of choices that also depend on risk attitudes. Often these involve strategic games, where the uncertain ways in which behavior of others deviate from standard predictions engenders a lottery for each player. Such uncertain deviations could be due to, for example, unobservable social preferences such as fairness or reciprocity. One example is the offer observed in ultimatum bargaining when the other player cannot be assumed to always accept a minuscule amount of money, and acceptable thresholds may be uncertain. Other examples include public goods contribution games where one does not know the extent of free riding of other players, trust games in which one does not know the likelihood that the other player will return some of the pie transferred to him, or Centipede games where one does not know when the other player will stop the game. Another source of uncertainty is the possibility that subjects make decisions with error, as predicted in Quantal Response Equilibria. Harrison (1989, 1990) and Harrison and Rutström (2008, Section 3.6) consider the use of controls for risk attitudes in bidding in first-price auctions.

In some cases, however, we simply want to elicit a preference from choices that do not depend on the choices made by others in a strategic sense, but which still depend on risk attitudes in a certain sense. An example due to Andersen et al. (2008a) is the elicitation of individual discount rates. In this case it is the concavity of the utility function that is important, and under EUT, this is synonymous with risk attitudes. Thus the risk aversion task is just a (convenient) vehicle to infer utility over deterministic outcomes. The implication is that we should combine a risk elicitation task with a time preference elicitation task and should use them jointly to infer discount rates over utility.

Defining Discount Rates in Terms of Utility

Assume that EUT holds for choices over risky alternatives and that discounting is exponential. A subject is indifferent between two income options M_t and $M_{t+\tau}$ if and only if

$$U(\omega + M_t) + (1/(1+\delta)^{\tau})U(\omega) = U(\omega) + (1/(1+\delta)^{\tau})U(\omega + M_{t+\tau}), \qquad (15.6)$$

where $U(\omega + M_{t+\tau})$ is the utility of monetary outcome M_t for delivery at time t plus some measure of background consumption ω, δ is the discount rate, τ is the horizon for delivery of the later monetary outcome at time $t + \tau$, and the utility function U is separable and stationary over time. The left-hand side of equation (15.6) is the sum of the discounted utilities of receiving the monetary outcome M_t at time t (in addition to background consumption) and receiving nothing extra at time $t + \tau$, and the right-hand side is the sum of the discounted utilities of receiving

nothing over background consumption at time t and the outcome $M_{t+\tau}$ (plus background consumption) at time $t+\tau$. Thus (15.6) is an indifference condition and δ is the discount rate that equalizes the present value of the *utility* of the two monetary outcomes M_t and $M_{t+\tau}$, after integration with an appropriate level of background consumption ω.

Most analyses of discounting models implicitly assume that the individual has a linear utility function,[20] so that (15.6) is instead written in the more familiar form

$$M_t = (1/(1+\delta)^{\tau})M_{t+\tau}, \tag{15.7}$$

where δ is the discount rate that makes the present value of the two monetary outcomes M_t and $M_{t+\tau}$ equal.

To state the obvious, (15.6) and (15.7) are not the same. As one relaxes the assumption that the decision maker has a linear utility function, it is apparent from Jensen's Inequality that the implied discount rate decreases if $U(M)$ is concave in M. Thus one cannot infer the level of the individual discount rate without knowing or assuming something about their utility function. This identification problem implies that discount rates cannot be estimated based on discount rate experiments with choices defined solely over time-dated money flows and that separate tasks to identify the extent of diminishing marginal utility must also be implemented.

Thus there is a clear implication from theory to experimental design: You need to know the nonlinearity of the utility function before you can *conceptually* define the discount rate. There is also a clear implication for the econometric method: You need to jointly estimate the parameters of the utility function and the discount rate, to ensure that sampling errors in one propagate correctly to sampling errors of the other. In other words, if we know the parameters of the utility function less precisely, due to small samples or poor parametric specifications, we have to use methods that reflect the effect of that imprecision on our estimates of discount rates.[21]

Andersen et al. (2008a) do this, and they infer discount rates for the adult Danish population that are well below those estimated in the previous literature that assumed linear utility functions, such as Harrison et al. (2002), who estimated annualized rates of 28.1% for the same target population. Allowing for concave utility, they obtain a point estimate of the discount rate of 10.1%, which is significantly lower than the estimate of 25.2% for the same sample assuming linear utility. This does more than simply verify that discount rates and diminishing marginal utility are mathematical substitutes in the sense that either of them have the effect of lowering the influence from future payoffs on present utility. It tells us that, for utility function coefficients that are reasonable from the standpoint of explaining choices in the lottery choice task,[22] the estimated discount rate takes on a value that is much more in line with what one would expect from market interest rates. To evaluate the statistical significance of adjusting for a concave utility function, one can test the hypothesis that the estimated discount rate assuming risk aversion is

the same as the discount rate estimated assuming linear utility functions. This null hypothesis is easily rejected. Thus, *allowing for diminishing marginal utility makes a significant difference to the elicited discount rates.*

The Need for Joint Estimation

We can write out the likelihood function for the choices that our subjects made and jointly estimate the risk parameter r in equation (15.1) and the discount rate δ. We use the same stochastic error specification as Holt and Laury (2002), and the contribution to the overall likelihood from the risk aversion responses is given by (15.5′).

A similar specification is employed for the discount rate choices. Equation (15.3) is replaced by the discounted utility of each of the two options, conditional on some assumed discount rate, and equation (15.4) is defined in terms of those discounted utilities instead of the expected utilities. The discounted utility of Option A is given by

$$PV_A = \frac{(\omega + M_A)^{(1-r)}}{1-r} + \frac{(1/(1+\delta)^{\tau})\omega^{(1-r)}}{1-r} \tag{15.8}$$

and the discounted utility of Option B is

$$PV_B = \frac{\omega^{(1-r)}}{1-r} + \frac{(1/(1+\delta)^{\tau})(\omega + M_B)^{(1-r)}}{1-r}, \tag{15.9}$$

where M_A and M_B are the monetary amounts in the choice tasks presented to subjects, illustrated in Table 15.2, and the utility function is assumed to be stationary over time.

An index of the difference between these present values, conditional on r and δ, can then be defined as

$$\nabla PV = \frac{PV_B^{1/\eta}}{PV_A^{1/\eta} + PV_B^{1/\eta}} \tag{15.10}$$

where η is a noise parameter for the discount rate choices, just as μ was a noise parameter for the risk aversion choices. It is not obvious that $\mu = \eta$, since these are cognitively different tasks. Our own priors are that the risk aversion tasks are harder, since they involve four outcomes compared to two outcomes in the discount rate tasks, so we would expect $\mu > \eta$. Error structures are things that one should always be agnostic about since they capture one's modeling ignorance, and we allow the error terms to differ between the risk and discount rate tasks.

Thus the likelihood of the discount rate responses, conditional on the EUT, CRRA, and exponential discounting specifications being true, depends on the

Table 15.2. Payoff Table for 6-Month Time Horizon Discount Rate Experiments

Payoff Alternative	**Payment Option A** (Pays Amount Below in 1 Month)	**Payment Option B** (Pays Amount Below in 7 Months)	**Annual Interest Rate** (AR, in %)	**Annual Effective Interest Rate** (AER, in %)	**Preferred Payment Option** (Circle A or B)	
1	3000 DKK	3038 DKK	2.5	2.52	A	B
2	3000 DKK	3075 DKK	5	5.09	A	B
3	3000 DKK	3114 DKK	7.5	7.71	A	B
4	3000 DKK	3152 DKK	10	10.38	A	B
5	3000 DKK	3190 DKK	12.5	13.1	A	B
6	3000 DKK	3229 DKK	15	15.87	A	B
7	3000 DKK	3268 DKK	17.5	18.68	A	B
8	3000 DKK	3308 DKK	20	21.55	A	B
9	3000 DKK	3347 DKK	22.5	24.47	A	B
10	3000 DKK	3387 DKK	25	27.44	A	B
11	3000 DKK	3427 DKK	27.5	30.47	A	B
12	3000 DKK	3467 DKK	30	33.55	A	B
13	3000 DKK	3507 DKK	32.5	36.68	A	B
14	3000 DKK	3548 DKK	35	39.87	A	B
15	3000 DKK	3589 DKK	37.5	43.11	A	B
16	3000 DKK	3630 DKK	40	46.41	A	B
17	3000 DKK	3671 DKK	42.5	49.77	A	B
18	3000 DKK	3713 DKK	45	53.18	A	B
19	3000 DKK	3755 DKK	47.5	56.65	A	B
20	3000 DKK	3797 DKK	50	60.18	A	B

estimates of r, δ, μ, and η, given the assumed value of ω and the observed choices.[23] If we ignore the responses that reflect indifference, the conditional log-likelihood is

$$\ln L(r, \delta, \mu, \eta; y, \omega, \mathbf{X}) = \sum_i \left[\left(\ln \Phi(\nabla PV) \times \mathbf{I}(y_i = 1) \right) + \left(\ln(1 - \Phi(\nabla PV)) \times \mathbf{I}(y_i = -1) \right) \right] \tag{15.11}$$

where $y_i = 1(-1)$ again denotes the choice of Option B (A) in discount rate task i, and $\mathbf{X}$ is a vector of individual characteristics.

The joint likelihood of the risk aversion and discount rate responses can then be written as

$$\ln L(r, \delta, \mu, \eta; y, \omega, \mathbf{X}) = \ln L^{RA} + \ln L^{DR}, \tag{15.12}$$

where L^{RA} is defined by (15.5′) and L^{DR} is defined by (15.11). This expression can then be maximized using standard numerical methods.

Lessons Learned

We draw together some methodological and practical lessons that we have learned from our research into risk and time preferences in Denmark. These are often reflections on our experience over many years in considering how theory, experimental design, and econometrics inform and constrain each other.

The Role of Artifactual Field Experiments

Harrison and List (2004) go to great lengths to point out that field experiments often entail many changes compared to traditional laboratory experiments. These changes involve sample composition, type of commodity, environment, information, stakes, literacy, and so on. When one observes differences in behavior in the field compared to the lab, which of these is driving that difference? Or, how do we know that there are not offsetting effects from different components of the field environment? For some inferential purposes we don't care about this sort of decomposition, but all too often we care deeply. The reason is that the "story" or "spin" that is put on the difference in behavior has to do with some structural component of the theory explaining behavior. For example, the claim in some quarters is that people exhibit more rational behavior in the field and that irrationality is primarily confined to the lab.[24] Or that people exhibit apparent altruism in the lab but rarely in the field. Or that "market interactions" fix all evils of irrationality. These are overstatements, but are not too far from the subplot of recent literature.

The critical role of "artifactual field experiments," as Harrison and List (2004) call them, is to take the simplest conceptual step over the bridge from the lab to

the field: Vary the composition of the sample from the convenience sample of university students. Although conceptually simple, the implementation is not always simple, particularly if one wants to generate representative samples of a large population as distinct from studying well-defined subsamples that mass conveniently at trade shows or other locations. Whether these are best characterized as being lab experiments or field experiments is not, to us, the real issue: The key thing is to see this type of experiment along a continuum taking one from the lab to the field, to better understand behavior.

To illustrate this path, consider the evaluation of risk attitudes in the field. Our field experiments in Denmark, reviewed in the section entitled "Risk Aversion," illustrate well the issues involved in taking the first step away from the lab. But to go further entails more than just "leaving the classroom" and recruiting outside of a university setting. In terms of sample composition, it means finding subjects who deal with that type of uncertainty in varying degrees and trying to measure the extent of their field experience with uncertainty. Moreover, it means developing stimuli that more closely match those that the subjects have previously experienced, so that they can use whatever heuristics they have developed for that commodity when making their choices. Finally, it means developing ways of communicating probabilities that correspond with language that is familiar to the subject. Thus, field experimentation in this case ultimately involves several simultaneous changes from the lab setting with respect to subject recruitment and the development of stimuli that match the field setting. Examples of studies that do these things, to varying degrees, are Harrison et al. (2007) and Fiore et al. (2009). In each case the changes were, by design, partial, so that one could better understand the effect on behavior. This is the essence of control that leads us to use experimental methods, after all.[25]

To see the concern with going into the field from the lab, consider the importance of "background risk" for the attitudes toward a specific "foreground risk" that are elicited. In many field settings it is simply not possible to artificially identify attitudes toward one risk source without worrying about how the subjects view that risk as being correlated with other risks. For example, mortality risks from alternative occupations tend to be highly correlated with morbidity risks: What doesn't kill you, sadly, often injures you. It is implausible to ask subjects their attitude toward one risk without some coherent explanation in the instructions as to why a higher or lower level of that risk would not be associated with a higher or lower risk of the other. In general, this will not be something that is amenable to field investigation in a controlled manner, although a few exceptions exist, as illustrated by Harrison et al. (2007).

In a similar vein, there is a huge literature on how one can use laboratory experiments to calibrate hypothetical field surveys for "hypothetical bias" in valuations (see Harrison (2006) for a review). In this case the value of the complementarity of field and lab is not due to concerns about the artifactual nature of the lab, but it is rather the artifactual nature of the field commodity that is causing problems.

For example, when someone asks you your willingness to pay $100 to reduce the risk of global warming, how should you interpret what you are actually buying? There is simply no way to run a naturally occurring field experiment in this case, or in any way that is free of major confounds. So evidence of hypothetical bias in many disparate private goods experiments can be used to condition the responses obtained in the field. For example, if women always respond identically in hypothetical and real lab experiments, and men state valuations that are always double what they would if it were real, then one surely has *some* basis for adjusting field hypothetical responses if one knows the sex of the respondent. The notion of calibration, introduced in this area by Blackburn et al. (1994), formalizes the statistical process of adjusting the field survey responses for the priors that one obtains in the lab environment.

The Contrived Debate Between Lab and Field Experiments

A corollary of the case for artifactual field experiments is the case for the complementarity of laboratory and field experiments. This theme was front and center in Harrison and List (2004), but appears to have been lost in some subsequent commentaries selling field experimental methodology. One illustration of this complementarity comes from our work and is motivated by the view that theory, experimental design, and econometrics are mutually dependent, resulting in numerous auxiliary hypotheses that can most efficiently be investigated in the lab. It is often unwieldy and inefficient to test procedures and treatments completely in the field, even if one would like to do so. The efficient mix is to identify the important treatments for application in the field, and they address the less important ones in the laboratory. Of course this involves some judgment about what is important, but we are often guided there by theory, previous evidence, the need for econometric identification, and, yes, one's nose for what sells in the journals. We have been careful in our own work to consider the balance between lab and field carefully at the outset. We expect that many others will do the same, so that the tendency to present field experiments as universally superior to lab experiments will organically shrink.

There are several examples of this complementarity from our work. In one case we used the laboratory to evaluate the performance of the iMPL elicitation procedure we assumed in the field. In the lab we could consider controlled comparisons to alternative methods and see what biases might have been generated by using the iMPL (see Andersen et al., (2006)). It would simply have been inefficient to carry all of those variants into the field.

In another example, we were concerned about the possibility of sample selection into experiments on the basis of risk attitudes: the so-called "randomization bias" much discussed in the broader experimental literature.[26] If subjects know that participation in experiments might entail randomization to treatment, and they have heterogeneous risk attitudes, then one would *a priori* expect to see less

risk averse subjects in experiments. But in experimental economics we offset that with a fixed, nonstochastic show-up fee, so what is the net effect? To be honest, we only thought of this after running our previous generation of field experiments, but could quickly evaluate the obvious experimental design in the lab with the same instruments (see Harrison et al., 2009). Finding evidence of sample selection, we now build the obvious design checks into the next generation of our field experiments.

Danes Are Like Plain Yogurt, Not Like Wines or Cheeses

We are well aware that results for Denmark, as important as they are for Danish policy, and perhaps also methodologically, might not readily transfer to other populations. In our case, much of our non-Danish work has involved developing countries, and the differences there can be dramatic. We would expect to see greater heterogeneity, greater instability, and perhaps even greater variety in the types of decision-making models employed. In this respect, however, we stress that the tools we have developed may be generally applied.

To take one important example, consider what one might mean by the "stability" of risk preferences over time. Does this mean that the unconditional estimate of RRA for each subject is the same over time, that the distribution for a given population stays the same even if individuals pop up at different parts of the distribution from time to time, or that the RRA or distribution are stable functions of observable states of nature that might change over time? In the latter case, where we think of states of nature as homely and intelligible things such as health, marital status, and family composition, it could be that preferences are a stable function of those states, but appear to be unstable when evaluated unconditionally. Using a longitudinal field experimental design, we examined exactly this question in Denmark (Anderson et al., 2008b). We found that preferences were generally stable, with some caveats, in virtually all three senses. But is this just a reflection of the "plain yogurt" of Danish culture, or something more general? Our personal priors may tell us one thing, but only data from new experiments can verify if this is true. But the point is that the longitudinal methodology, along with questionnaires on states of nature, is a general approach applicable beyond the specific population of Denmark.

Non-EUT Models of Risky Choice and Non-Exponential Models of Discounting: *Festine Lente*

We have serious doubts about the generality and robustness of some of the empirical claims of the literature with respect to risk and time preferences. Much of the empirical evidence has been obtained by staring at "patterns" of choices rather than estimating structural parameters and testing for statistical significance. It is quite possible for there to be statistically significant differences in patterns of choices, according to some unconditional semiparametric test, but for that to be consistent

with a wide range of underlying structural models. This is particularly true when one augments those models with alternative "behavioral error" stories. To take one simple example, the violations of first-order stochastic dominance that motivated the shift from original prospect theory to cumulative prospect theory can, to some extent, be accounted for by certain error specifications. One might not want to do that, and we are not advocating it as a general econometric policy, but the point is that inferences about patterns are not the same as inferences about the latent structural parameters.

Another issue with much of the received evidence for violations of EUT or exponential discounting is that it has been drawn from convenience samples in laboratory experiments. Relatively little evidence has been collected from field experiments with a broader sample from the population, using comparable instruments and/or instruments that arise more naturally in the decision-making environments of the subjects. We are not denying that "students are people too," just noting that they have a distinct set of demographic characteristics that can matter significantly for policy inferences (e.g., Andersen et al., 2010) and that all exhibit one characteristic that might reflect sample selection on unobservables of relevance for the experimental task at hand (viz., their presence at a college or university).

Finally, our work leads us to question some of the inferential assumptions of previous tests. Mixture specifications, in which one allows two or more data-generating processes to explain observed behavior, show clear evidence that behavior is not wholly explained by any one of the popular models. Andersen et al. (2008a, Section 3.D) consider a mixture specification of exponential and hyperbolic discounting and find that 72% of the choices are better characterized as exponential.[27] This estimate of the mixing probability is statistically significantly different from 0% or 50%. Similarly, Harrison and Rutström (2009) find roughly equal support for EUT and Prospect Theory in a lab setting; Harrison et al. (2010) find roughly equal support for EUT and Rank-Dependent Utility models in artifactual field experiments in India, Ethiopia, and Uganda; and Coller et al. (2012) find roughly equal support for exponential and quasi-hyperbolic discounting in the laboratory.

The key insight from mixture specifications is to simply change the question that is posed to the data. Previous econometric analyses have posed a proper question: If one and only one data-generating process is to account for these data, what are the estimated parameter values and do they support a nonstandard specification? The simplest, finite mixture specification changes this to the following: If two data-generating processes are allowed to account for the data, what fraction is attributable to each, and what are the estimated parameter values? Nevertheless, one can imagine someone still wanting to ask the former question, if they just wanted one "best" model. But that question is also seen to constrain evidence of heterogeneity of decision-making processes, and we prefer to avoid that when we can.[28] There are fascinating issues with the specific implementation and interpretation of mixture models, but those are not germane to the main insight they provide.[29]

Finally, there are often simple issues of functional specification which can only be explored with structural models. For example, what happens when subjects bring an unobserved "homegrown reference point" to the experimental task and the analyst tries to infer measures of loss aversion? In other words, what if the subject rationally expects to get more than just the show-up fee? The answer is that one gets extremely sensitive estimates of loss aversion depending on what one assumes, not too surprisingly (e.g., Harrison and Rutström, 2008, Section 3.2.3). This is likely to be a more serious issue in the field, due to a greater diversity in homegrown reference points. As another example, some tests of EUT rest on the assumption that subjects use a very restrictive functional form. The tests of myopic loss aversion offered in Gneezy and Potters (1997) rest on assuming that CRRA and their "violations of EUT" can be accounted for simply with an expo-power specification that allows varying RRA with prize level (e.g., Harrison and Rutström, 2008, Section 3.7). So the initial evidence does show a violation of CRRA, but this is not something that one ought to get too excited about.

We do not want to overstate the case for standard specifications. We do find some evidence for *some* subjects to behave differently than typically assumed in EUT and exponential discounting specifications, in *some* tasks. It is just that the evidence is hardly as monolithic as many claim.

Estimation, Not Direct Elicitation

When we began our research, we were focused on designing instruments that could directly provide more precise estimates of risk attitudes and temporal preferences. Our work on the iMPL, discussed earlier, was directly solely at that objective: refining down the interval within which we had captured the true, latent risk attitude or discount rate. We certainly believe that those procedures accomplish that goal, but our focus quickly moved away from designing a better mousetrap to learning more about the mouse itself. That is, we discovered that the questions we wanted to ask demanded that we employ structural estimation, if for no other reason than to be able to condition inferences from the discount rate task with estimates of the utility function (from the risk aversion task). This need for joint estimation, along with full information maximum likelihood recognition that errors in estimation of the utility function *should* propagate into errors in estimation of discount rates, as a matter of theory as well as econometric logic, is much more general than these examples. Inferences about bidding in auctions, about subjective beliefs elicited from a proper scoring rule, about social preferences, and about behavior in games with payoffs defined over utility all require that one say something about utility functions unless one is working on special cases. Indeed, there is mounting evidence that the inferences change dramatically when one allows for nonlinear utility functions.[30] The days of designing tasks to elicit exactly what one wants in one task are long over.

We also see no tension whatsoever between the interest in using experiments to generate estimates of latent structural parameters and the interest in using

experiments to test comparative static propositions from theory. The assertion that economic theory "works" when we do the latter, but not the former, is just that, an assertion.[31] If we are to understand why certain comparative static outcomes occur, or do not, we need to know what moving parts of the underlying theory are misspecified, as well as if there need to be more "moving parts" added to the theory. Quite apart from the value of knowing this from a descriptive point of view, normative inferences demand knowledge of this kind if they are to be more than black box behavioral assertions.

Virtual Experiments as a Smooth Bridge Between the Lab and the Field

It is now wellknown and accepted that behavior is sensitive to the cognitive constraints of participants. It has been recognized for some time that field referents and cues are essential elements in the decision process, and can serve to overcome such constraints (Ortmann and Gigerenzer, 1997), even if there are many who point to "frames" as the source of misbehavior from the perspective of traditional economic theory (Kahneman and Tversky, 2000). The concept of "ecological rationality" captures the essential idea of those who see heuristics as potentially valuable decision tools (Gigerenzer and Todd, 1999; Smith, 2003). According to this view, cognition has evolved within specific decision environments. If that evolution is driven by ecological fitness, then the resulting cognitive structures, such as decision heuristics, are efficient and accurate *within these environments*. But they may often fail when applied to new environments.

At least two other research programs develop similar views. Glimcher (2003) describes a research program, following Marr (1982), that argues for understanding human decision making as a function of a complete biological system rather than as a collection of mechanisms. As a biological system, he views decision-making functions as having evolved to be fit for specific environments. Clark (1997) sees cognition as extended outside not just the brain but the entire human body, defining it in terms of all the tools used in the cognitive process, both internal and external to the body. Field cues can be considered external aspects of such a process. Behavioral economists are paying attention to these research programs and what they imply for the understanding of the interactions between the decision maker and his environment. For our purposes here, it means we have to pay careful attention to the role of experiential learning in the presence of specific field cues and how this influences decisions.

The acceptance of the role of field cues in cognition provides arguments in favor of field rather than lab experiments (Harrison and List, 2004). Where else than in field experiments can you study decision makers in their natural environment using field cues that they have come to depend on? We actually challenge this view, if it is taken to argue that the laboratory environment is *necessarily* unreliable (Levitt and List, 2007). While it is true that lab experiments traditionally use

artifactual and stylized tasks that are free of field cues, in order to generate the type of control that is seen as essential to hypothesis testing, field experiments have other weaknesses that a priori are equally important to recognize (Harrison, 2005). Most importantly, the ability to implement necessary controls on experimental conditions in the field is much more limited than in the lab, as is the ability to implement many counterfactual scenarios. In addition, recruitment is often done in such a way that it is difficult to avoid and control for sample selection effects; indeed, in many instances the natural process of selection provides the very treatment of interest (e.g., Harrison and List, 2008). However, this means that one must take the sample with all of the unobservables that it might have selected, and just assume that they did not interact with the behavior being measured. Finally, the cost of generating observational data can be quite significant in the field, at least in comparison to the lab.

For all these reasons, we again see lab and field experiments as complementary, a persistent theme of Harrison and List (2004). A proper understanding of decision making requires the use of both. While lab experiments are better at generating internal validity, imposing the controlled conditions necessary for hypothesis testing, field experiments are better at generating external validity, including the natural field cues.

Fiore et al. (2009) propose a new experimental environment, the Virtual Experiment (VX), that has the potential of generating both the internal validity of lab experiments and the external validity of field experiments. A VX is an experiment set in a controlled lab-like environment, using either typical lab or field participants, that generates synthetic field cues using Virtual Reality (VR) technology. The experiment can be taken to typical field samples, such as experts in some decision domain, or to typical lab samples, such as student participants. The VX environment can generate internal validity since it is able to closely mimic explicit and implicit assumptions of theoretical models and thus provide tight tests of theory; it is also able to replicate conditions in past experiments for robustness tests of auxiliary assumptions or empirically generated hypotheses. The VX environment can generate external validity because observations can be made in an environment with cues mimicking those occurring in the field. In addition, any dynamic scenarios can be presented in a realistic and physically consistent manner, making the interaction seem natural for the participant. Thus the VX builds a bridge between the lab and the field, allowing the researcher to smoothly go from one to the other and see what features of each change behavior. VX is a methodological frontier enabling new levels of understanding via integration of laboratory and field research in ways not previously possible. Echoing calls by others for such an integration, we argue that "research must be conducted in various settings, ranging from the artificial laboratory, through the naturalistic laboratory, to the natural environment itself" (Hoffman and Deffenbacher, 1993, p. 343).

The potential applications for VX are numerous. Apart from simulating actual policy scenarios, such as the wild fire prevention policies investigated by

Fiore et al. (2009), it can also be used to mimic environments assumed in a number of field data analyses. For example, popular ways of estimating valuations for environmental goods include the Travel Cost Method (TCM), the Hedonic Pricing Method (HPM), and the Stated Choice Method (SCM). To mimic TCM, the simulation can present participants with different travel alternatives and observe which ones are chosen under different naturalistic conditions. To mimic HPM, the simulation can present participants with different real estate options and observe purchasing behavior, or simply observe pricing behavior for alternative options (Castronova, 2004). Finally, to mimic SCM, participants can experience the different options they are to choose from through naturalistic simulation. For all of these types of scenarios, some of the most powerful applications of VX will involve continuous representations of dynamically generated effects of policy changes. Visualizing and experiencing long-term effects correctly should improve short-run decisions with long-run consequences.

In the application to wild fire prevention policies, we use actual choices by subjects that bear real economic consequences from those choices. Participants are presented with two options: One simply continues the present fire prevention policies, and the other increases the use of prescribed burns. Participants get to experience two fire seasons under each policy and are then asked to make a choice between them. The scenario that simulates the continuation of the present fire prevention policies will realistically generate fires that cause more damage on average and that also vary substantially in intensity. This option therefore presents the participant with a risky gamble with low expected value. The alternative option presents a relatively safe gamble with a higher expected value, but there will be a nonstochastic cost involved in implementing the expansion of prescribed burns. It is possible in VX to set the payoff parameters in such a way that one can estimate Willingness to Pay (WTP) for the burn expansion option that is informative to actual fire policy. These values of WTP could then be compared to those generated through a popular Contingent Valuation Method to test the hypothesis that they should be different. Alternatively, it is possible to manipulate the payoff parameters in such a way that one estimates parameters of choice models such as risk attitudes, loss aversion, and probability weights.

In summary, we see the use of VX as a new tool in experimental economics, with an emphasis on the methodological issues involved in bridging the gap between lab and field experiments.

Conclusions

We love doing experimental economics because it provides a unique ability to directly and explicitly confront theorists and econometricians with real behavior. There is no hiding behind theoretical models that claim to be operationally meaningful but never really come close, and there is no hiding behind proxies for

variables of theoretical interest when the experimental design is developed rigorously. This work requires modest extensions of the tools used by experimental economists, to (a) develop designs that employ several tasks to identify all of the moving parts of the theoretical machine and (b) write out the likelihood of observed behavior using the structural model being tested, but these are feasible and well-understood components of modern experimental economics.

It is then natural to combine laboratory and field experiments in this enterprise. When we think of new areas of our own research, such as the elicitation of subjective beliefs and the measurement of aversion to uncertainty, we would never think of first developing experimental designs in the field. Actually, this is also true of older areas of research which we believe to be often intractably confounded in the field, such as "social preferences," "trust," and "loss aversion," but our premise there is apparently not widely shared. In any event, the efficient frontier for this production process demands that we combine laboratory and field environments. They can be substituted to varying degrees, of course, but it is difficult for us to imagine an area of enquiry that would not benefit from some inputs of both.

NOTES

Steffen Andersen and Melonie Sullivan have made significant contributions to the research discussed here. We thank the U.S. National Science Foundation for research support under grants NSF/HSD 0527675 and NSF/SES 0616746, and we are grateful to the Danish Social Science Research Council for research support under projects 24-02-0124 and 275-08-0289.

1. There is a growing literature of experiments performed outside of university research laboratories, building on the pioneering work of Peter Bohm over many years, starting in the 1970s. Dufwenberg and Harrison (2008, p. 214ff) provide a posthumous appreciation of his motivation: "Peter was drawn to conduct field experiments long before laboratory experiments had become a staple in the methodological arsenal of economists. Just as some experimentalists do not comprehend why one would ask questions with no real economic consequences, or care too much about the responses to such questions, Peter began doing field experiments simply because they answered the questions he was interested in. He did not come to field experiments because of any frustration with lab experiments or [because of] any long methodological angst about laboratory experiments: [I]t was just obvious to him that experiments needed field referents to be interesting. He later became interested in the methodological differences between laboratory and field experiments, well after his own pioneering contributions to the later had been published." Due to the great variety of such experiments with respect to procedures, contexts and participant pools there has been a refinement of the field-lab terminology to include modifiers such as "artifactual." We will restrict our discussions to two kinds of experiments only: (a) the traditional research laboratory using convenient and low-cost student samples and (b) the artifactual field experiment that employs samples from populations not restricted to students. In these latter experiments the tasks are similar to those presented to students but often have to be adjusted to the perceptual and conceptual needs of the subject pool. Here we will simply use the label "lab" when referring to

experiments we conduct on student samples and "field" to those conducted on samples from more heterogeneous field populations.

2. For example, see Desvousges et al. (1999). The limitation on information can derive from the inherent difficulty of modeling behavioral or physical relationships, from the short time-frame over which the model has to be developed and applied, or both.

3. Revenue neutrality is defined in terms of real government revenue and does not imply welfare neutrality.

4. For example, if the empirical distribution of the elasticity of substitution is specified to be normal with mean 1.3 and standard deviation 0.4, 95% of the random draws will be within $\pm 1.96 \times 0.4$ of the mean. Thus one would rarely see this elasticity take on values greater than 3 or 4 in the course of these random draws.

5. Defined by the 25th and 75th percentiles, this range represents 50% of the observations around the median.

6. The manner in which these sidepayments are computed is explained in Harrison et al. (2002). It corresponds to a stylized version of the type of political balancing act one often encounters behind the scenes in the design of a public policy such as this.

7. For example, if the elasticity of demand for a product with a large initial indirect tax is higher than the default elasticity, households can substitute toward that product more readily and enjoy a higher real income for any given factor income.

8. We control for county and the recruitment wave to which the subject responded in our statistical analysis of sample selection. Response rates were higher in the greater Copenhagen area compared to the rest of the country. The experiments were conducted under the auspices of the Ministry of Economic and Business Affairs, and people living outside of the greater Copenhagen area may be suspicious of government employees and therefore less likely to respond to our letter of invitation.

9. The first person suffered from dementia and could not remember the instructions; the second person was a 76-year-old woman who was not able to control the mouse and eventually gave up; the third person had just won a world championship in sailing and was too busy with media interviews to stay for two hours; and the fourth person was sent home because they arrived after the instructions had begun and we had already included one unexpected "walk-in" to fill their position.

10. Certain events might have plausibly triggered some of the no-shows: for example, three men did not turn up on June 11, 2003, but that was the night that the Danish national soccer team played a qualifying game for the European championships against Luxembourg that was not scheduled when we picked session dates.

11. We are implicitly assuming that the utility function of the subject is only defined over the prizes of the experimental task. We discuss this assumption below.

12. That is, if someone decides at some stage to switch from option A to option B between probability 0.4 and 0.5, the next stage of an iMPL would then prompt the subject to make more choices within this interval for probabilities form 0.40 to 0.50 increasing by 0.01 on each row. The computer implementation of the iMPL restricts the number of stages to ensure that the intervals exceed some a priori cognitive threshold (e.g., probability increments of 0.01).

13. In our experience, subjects are suspicious of randomization generated by computers. Given the propensity of many experimenters in other disciplines to engage in deception, we avoid computer randomization whenever feasible.

14. In an important respect, joint estimation can be viewed as Full Information Maximum Likelihood (FIML) since it uses the entire set of structural equations from theory to define the overall likelihood.

15. Our treatment of indifferent responses uses the specification developed by Papke and Wooldridge (1996, equation 5, p. 621) for fractional dependent variables. Alternatively, one could follow Hey and Orme (1994, p. 1302) and introduce a new parameter τ to capture the idea that certain subjects state indifference when the latent index showing how much they prefer one lottery over another falls below some threshold τ in absolute value. This is a natural assumption to make, particularly for the experiments they ran in which the subjects were told that expressions of indifference would be resolved by the experimenter, but not told how the experimenter would do that (p. 1295, footnote 4). It adds one more parameter to estimate, but for good cause.

16. Clustering commonly arises in national field surveys from the fact that physically proximate households are often sampled to save time and money, but it can also arise from more homely sampling procedures. For example, Williams (2000, p. 645) notes that it could arise from dental studies that "collect data on each tooth surface for each of several teeth from a set of patients" or "repeated measurements or recurrent events observed on the same person." The procedures for allowing for clustering allow heteroskedasticity between and within clusters, as well as autocorrelation within clusters. They are closely related to the "generalized estimating equations" approach to panel estimation in epidemiology (see Liang and Zeger (1986)), and they generalize the "robust standard errors" approach popular in econometrics (see Rogers (1993)). Wooldridge (2003) reviews some issues in the use of clustering for panel effects, noting that significant inferential problems may arise with small numbers of panels.

17. Exactly the same insight in a strategic context leads one from Nash Equilibria to Quantal Response Equilibria, if one reinterprets Figures 15.2 and 15.3, respectively, in terms of best-response functions defined over expected (utility) payoffs from two strategies. The only difference in the maximum likelihood specification is that the equilibrium condition jointly constrains the likelihood of observing certain choices by two or more players.

18. Some specifications place the error at the final choice between one lottery or after the subject has decided which one has the higher expected utility; some place the error earlier, on the comparison of preferences leading to the choice; and some place the error even earlier, on the determination of the expected utility of each lottery.

19. Reiley (2015) asks, "Are we measuring utility over income in the experiment? Utility over annual income? Utility over different wealth levels? Though not explicitly assumed in this work, researchers often assume implicitly that different types of risk preferences are the same. [. . .] I want us to be more aware of these assumptions we're making." Our discussion demonstrates that the literature is in fact quite explicit about these issues and has already progressed to testing different assumptions.

20. See Keller and Strazzera (2002, p. 148) and Frederick et al. (2002, p. 381ff) for an explicit statement of this assumption, which is often implicit in applied work. We occasionally refer to risk aversion and concavity of the utility function interchangeably, but it is concavity that is central (the two can differ for non-EUT specifications).

21. It is true that one must rely on structural assumptions about the form of utility functions, probability weighting functions, and discounting functions, in order to draw inferences. These assumptions can be tested, and have been, against more flexible versions and even nonparametric versions (e.g., Harrison and Rutström, 2008, p. 78–79). A similar debate rages with respect to structural assumptions about error specifications, as illustrated by the charming title of the book by Angrist and Pischke, (2009), *Mostly Harmless Econometrics*. But it is an illusion, popular in some quarters, that one can safely dispense with all structural assumptions and draw inferences: see Keane (2010) and Leamer (2010) for spirited assaults on that theology.

22. It is important to recognize that it is not risk attitudes per se that are important for identifying the discount rate, but instead estimates of the extent of diminishing marginal utility which allow one to condition inferences about discount rates defined over utility from observed choices over time-dated money flows. It just happens that inferring the curvature of the utility function is straightforward from choices over risky lottery tasks. For example, Reiley (2015) notes that it is ". . . an important observation that consistent estimates of time preferences depend on measurements of individuals' risk preferences." This suggests that the connection between risk preferences and time preferences is the result of some black behavioral box. Instead, it is the direct reflection of getting the theory right (i.e., that discount rates are *defined* in terms of utility flows) and then designing the experiment to estimate the conceptually right thing that determines whether one does this in the laboratory or the field. For example, what if risk preferences were best characterized using a dual theory representation in which utility functions were linear but decision makers exhibited pessimism in probability weighting? In that case, risk preferences per se would have no effect on inferences about discount rates, theoretically or empirically.

23. For simplicity we are implicitly assuming that the λ parameter from Andersen et al. (2008a) is equal to 1. This means that delayed experimental income is spent in one day.

24. And hence that we can safely dismiss the messy claims of behaviorists as artifacts of the lab. We might agree with this conclusion even if we do not agree with this argument for it (Harrison, 2010).

25. Reiley (2015) comments that control is not *always* a good thing and that the essence of a field experiment entails *some* lack of control in relation to laboratory experiments. We agree, as qualified, but this issue is more nuanced than saying that loss of control and internal validity is the price that one has to pay for external validity. One concern is that this view can be used as an excuse by field experimenters to get on with telling a story that seems plausible behaviorally, but that glosses confounds that could have been relatively easily controlled: Harrison (2005) offers examples. Another concern is that the artifactual devices used to try to ensure control in the laboratory, or the field, might themselves result in less control in terms of the inferences one intends to draw: Harrison and List (2004, Section 5) provide several examples, such as the widespread use of abstraction from field referents in the instructions used in laboratory experiments.

26. Since the use of randomized control trials has become popular in field experiments conducted in developing countries (Banerjee and Duflo, 2009), it is worth noting that they originated in a remarkable paper by Peirce and Jastrow (1885), although of course Fisher popularized their use. This early study was in response to the use of data—generated by Fechner, using himself as subject and experimenter—on the ability to discriminate between psycho-physical sensations from stimuli. These data, and Fechner's theorizing about it, evolved into what we referred to earlier as "the Fechner error." It is ironic that what we now see as an important insight in the structural modeling of latent behavior generated the statistical method that is now viewed by some as a vehicle for nonstructural, atheoretic data gathering.

27. Those experiments employed a Front-End Delay on payments of 30 days and were not designed to test Quasi-Hyperbolic specifications. The latest series of field experiments in Denmark, completed late 2009, are designed to test that specification *inter alia*.

28. Growing interest in the application of mixture specifications is a direct and constructive response to the issue posed by Reiley (2015): "Even though we go to great lengths to estimate accurate standard errors *given the model*, we don't ever inflate our

standard errors to the extent that we are uncertain *about the model itself* [. . .]." The literature has, in fact, been doing this for some time now, although it does not at all follow that the standard errors on specific parameters get larger as we allow for mixtures. In fact, there are notable examples where the estimates are significantly "better" by some measures, such as the loss aversion and probability weighting parameters of prospect theory (e.g., Harrison and Rutström, 2008, p. 95–100).

29. For example, does one constrain individuals or task types to be associated with just one data-generating process, or allow each choice to come from either? Does one consider more than two types of processes, using some specification rule to decide if there are 2, or 3, or more? Does one specify general models for each data-generating process and see if one of them collapses to a special case, or just specify the competing alternatives explicitly from the outset? How does one check for global maximum likelihood estimates in an environment that might generate multimodal likelihood functions "naturally"? Harrison and Rutström (2009) discuss these issues, and they point to the older literature.

30. There is no intended slight of models of decision making under risk that focus on things other than the linearity of the utility function; this point is quite general.

31. Reiley (2015) endorses this assertion and then makes the separate, but valid, point that one test of the estimates of structural models would be whether they can predict out of context and domain.

REFERENCES

Abdellaoui, M., C. Barrios, and P. P. Wakker. 2007. Reconciling Introspective Utility with Revealed Preference: Experimental Arguments Based on Prospect Theory. *Journal of Econometrics* **138**:356–378.

Andersen, S., J. C. Cox, G. W. Harrison, M. I. Lau, E. E. Rutström, and V. Sadiraj. 2011. Asset Integration and Attitudes to Risk: Theory and Evidence. Working Paper 2011-07, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2006. Elicitation Using Multiple Price Lists. *Experimental Economics* **9**(4): 383–405.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2008a. Eliciting Risk and Time Preferences. *Econometrica* **76**(3): 583–619.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2008b. Lost in State Space: Are Preferences Stable? *International Economic Review* **49**(3):1091–1112.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2008c. Risk Aversion in Game Shows. In *Risk Aversion in Experiments*, eds. J. C. Cox and G. W. Harrison. Greenwich, CT: JAI Press.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström. 2010. Preference Heterogeneity in Experiments: Comparing the Lab and Field. *Journal of Economic Behavior & Organization* **74**:209–224.

Angrist, J. D. and J. S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Banerjee, A. V. and E. Duflo. 2009. The Experimental Approach to Development Economics. *Annual Review of Economics* **1**:151–178.

Barr, A. and T. Packard. 2002. Revealed Preference and Self Insurance: Can We Learn from the Self Employed in Chile? *Policy Research Working Paper #2754*, World Bank, Washington, DC.

Blackburn, M., G. W. Harrison, and E. Rutström. 1994. Statistical Bias Functions and Informative Hypothetical Surveys. *American Journal of Agricultural Economics* **76**(5):1084–1088.

Camerer, C. and T. H. Ho. 1994. Violations of the Betweenness Axiom and Nonlinearity in Probability. *Journal of Risk & Uncertainty* **8**:167–196.

Castronova, E. 2004. The Price of Bodies: A Hedonic Pricing Model of Avatar Attributes in a Synthetic World. *Kyklos* **57**(2):173–196.

Chambers, R. G. and J. Quiggin. 2000. *Uncertainty, Production, Choice, and Agency: The State-Contingent Approach*. New York: Cambridge University Press.

Clark, A. 1997. *Being There*. Cambridge, MA: MIT Press.

Coller, M., G. W. Harrison, and E. E. Rutström. 2012. Latent Process Heterogeneity in Discounting Behavior. *Oxford Economic Papers* **64**(2):375–391.

Cox, J. C. and V. Sadiraj. 2006. Small- and Large-Stakes Risk Aversion: Implications of Concavity Calibration for Decision Theory. *Games & Economic Behavior* **56**:45–60.

Cox, J. C. and V. Sadiraj. 2008. Risky Decisions in the Large and in the Small: Theory and Experiment. In *Risk Aversion in Experiments*, eds. J. C. Cox and G. W. Harrison. Greenwich, CT: JAI Press.

Desvousges, W. H., F. R. Johnson, and H. S. Banzhaf. 1999. *Environmental Policy Analysis with Limited Information: Principles and Applications of the Transfer Method*. New York: Elgar.

Dufwenberg, M. and G. W. Harrison. 2008. Peter Bohm: Father of Field Experiments. *Experimental Economics* **11**(3):213–220.

Fiore, S. M., G. W. Harrison, C. E. Hughes, and E. E. Rutström. 2009. Virtual Experiments and Environmental Policy. *Journal of Environmental Economics & Management* **57**(1):65–86.

Frederick, S., G. Loewenstein, and T. O'Donoghue. 2002. Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature* **40**(2):351–401.

Gigerenzer, G. and P. M. Todd (eds.). 1999. *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Glimcher, P. 2003. *Decisions, Uncertainty and the Brain*. Cambridge, MA: MIT Press.

Gneezy, U. and J. Potters. 1997. An Experiment on Risk Taking and Evaluation Periods. *Quarterly Journal of Economics* **112**:631–645.

Grether, D. M. and C. R. Plott. 1979. Economic Theory of Choice and the Preference Reversal Phenomenon. *American Economic Review* **69**(4):623–648.

Hansson, B. 1988. Risk Aversion as a Problem of Conjoint Measurement. In *Decision, Probability, and Utility*, eds. P. Gardenfors and N. E. Sahlin. New York: Cambridge University Press.

Harrison, G. W. 1989. Theory and Misbehavior of First-Price Auctions. *American Economic Review* **79**:749–762.

Harrison, G. W. 1990. Risk Attitudes in First-Price Auction Experiments: A Bayesian Analysis. *Review of Economics & Statistics* **72**:541–546.

Harrison, G. W. 2005. Field Experiments and Control. In *Field Experiments in Economics*, eds. J. Carpenter, G. W. Harrison and J. A. List. Greenwich, CT: JAI Press.

Harrison, G. W. 2006. Experimental Evidence on Alternative Environmental Valuation Methods. *Environmental and Resource Economics* **34**:125–162.

Harrison, G. W. 2010. The Behavioral Counter-Revolution. *Journal of Economic Behavior & Organization* **73**:49–57.
Harrison, G. W., S. J. Humphrey, and A. Verschoor. 2010. Choice Under Uncertainty: Evidence from Ethiopia, India and Uganda. *Economic Journal* **120**:80–104.
Harrison, G. W., S. E. Jensen, L. Pedersen, and T. F. Rutherford. 2000. *Using Dynamic General Equilibrium Models for Policy Analysis*. Amsterdam: Elsevier.
Harrison, G. W., J. Jensen, M. I. Lau, and T. F. Rutherford. 2002. Policy Reform Without Tears. In *Policy Evaluation with Computable General Equilibrium Models*, eds. A. Fossati and W. Weigard. New York: Routledge.
Harrison, G. W., M. I. Lau, and E. Rutström. 2007. Estimating Risk Attitudes in Denmark: A Field Experiment. *Scandinavian Journal of Economics* **109**(2):341–368.
Harrison, G. W., M. I. Lau, and E. E. Rutström. 2009. Risk Attitudes, Randomization to Treatment, and Self-Selection Into Experiments. *Journal of Economic Behavior and Organization* **70**(3):498–507.
Harrison, G. W., M. I. Lau, E. E. Rutström, and M. B. Sullivan. 2005. Eliciting Risk and Time Preferences Using Field Experiments: Some Methodological Issues. In *Field Experiments in Economics*, eds. J. Carpenter, G. W. Harrison, and J. A. List. Greenwich, CT: JAI Press.
Harrison, G. W., M. I. Lau, and M. B. Williams. 2002. Estimating Individual Discount Rates for Denmark: A Field Experiment. *American Economic Review* **92**(5):1606–1617.
Harrison, G. W. and J. A. List. 2004. Field Experiments. *Journal of Economic Literature* **42**(4):1013–1059.
Harrison, G. W. and J. A. List. 2008. Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner's Curse. *Economic Journal* **118**:822–843.
Harrison, G. W., J. A. List, and C. Towe. 2007. Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion. *Econometrica* **75**(2):433–458.
Harrison, G. W., T. F. Rutherford, and D. G. Tarr. 2003. Trade Liberalization, Poverty and Efficient Equity. *Journal of Development Economics* **71**:97–128.
Harrison, G. W., T. F. Rutherford, D. G. Tarr, and A. Gurgel. 2004. Trade Policy and Poverty Reduction in Brazil. *World Bank Economic Review* **18**(3):289–317.
Harrison, G. W. and E. E. Rutström. 2008. Risk Aversion in the Laboratory. In *Risk Aversion in Experiments*, eds. J. C. Cox and G. W. Harrison. Greenwich, CT: JAI Press.
Harrison, G. W. and E. E. Rutström. 2009. Expected Utility and Prospect Theory: One Wedding and A Decent Funeral. *Experimental Economics* **12**(2):133–158.
Harrison, G. W. and H. D. Vinod. 1992. The Sensitivity Analysis of Applied General Equilibrium Models: Completely Randomized Factorial Sampling Designs. *Review of Economics and Statistics* **74**:357–362.
Hey, J. D. and C. Orme. 1994. Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica* **62**(6):1291–1326.
Hirshleifer, J. and J. G. Riley. 1992. *The Analytics of Uncertainty and Information*. New York: Cambridge University Press.
Hoffman, R. R. and K. A. Deffenbacher. 1993. An Analysis of the Relations of Basic and Applied Science. *Ecological Psychology* **5**:315–352.
Holt, C. A. and S. K. Laury. 2002. Risk Aversion and Incentive Effects. *American Economic Review* **92**(5):1644–1655.

Kahneman, D. and A. Tversky (eds.). 2000. *Choices, Values and Frames*. New York: Cambridge University Press.

Keane, M. P. 2010. Structural vs. Atheoretic Approaches to Econometrics. *Journal of Econometrics* **156**:3–20.

Keller, L. R. and E. Strazzera. 2002. Examining Predictive Accuracy Among Discounting Models. *Journal of Risk and Uncertainty* **24**(2):143–160.

Lau, M. I. 2000. Assessing Tax Reforms When Human Capital Is Endogenous. In *Using Dynamic General Equilibrium Models for Policy Analysis*, eds. G. W. Harrison, S. E. H. Jensen, L. H. Pedersen, and T. F. Rutherford. Amsterdam: North Holland.

Leamer, E. E. 2010. Tantalus on the Road to Asymptotia. *Journal of Economic Perspectives* **24**(2):31–46.

Levitt, S. D. and J. A. List. 2007. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World. *Journal of Economic Perspectives* **21**(2):153–174.

Liang, K. Y. and S. L. Zeger. 1986. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73**:13–22.

List, J. A. 2005. Scientific Numerology, Preference Anomalies, and Environmental Policymaking. *Environmental & Resource Economics* **32**:35–53.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman & Company.

Miller, L., E. David, and J. T. Lanzetta. 1969. Choice Among Equal Expected Value Alternatives: Sequential Effects of Winning Probability Level on Risk Preferences. *Journal of Experimental Psychology* **79**(3):419–423.

Ortmann, A. and G. Gigerenzer. 1997. Reasoning in Economics and Psychology: Why Social Context Matters. *Journal of Institutional and Theoretical Economics* **153**(4):700–710.

Papke, L. E. and J. M. Wooldridge. 1996. Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics* **11**:619–632.

Peirce, C. S. and J. Jastrow. 1885. On Small Differences of Sensation. *Memoirs of the National Academy of Sciences for 1884* **3**:75–83.

Rabin, M. 2000. Risk Aversion and Expected Utility Theory: A Calibration Theorem. *Econometrica* **68**:1281–1292.

Reiley, D. 2015. The Lab and the Field: Empirical and Experimental Economics. In *Handbook of Experimental Economics Methodology*, eds. G. Frechette and A. Schotter. New York: Oxford University Press.

Rogers, W. H. 1993. Regression Standard Errors in Clustered Samples. *Stata Technical Bulletin* **13**:19–23.

Safra, Z. and U. Segal. 2008. Calibration Results for Non-Expected Utility Theories. *Econometrica* **76**(5):1143–1166.

Saha, A. 1993. Expo-Power Utility: A Flexible Form for Absolute and Relative Risk Aversion. *American Journal of Agricultural Economics* **75**(4):905–913.

Schubert, R., M. Brown, M. Gysler, and H. W. Brachinger. 1999. Financial Decision-Making: Are Women Really More Risk-Averse? *American Economic Review (Papers & Proceedings)* **89**(2):381–385.

Smith, V. L. 2003. Constructivist and Ecological Rationality in Economics. *American Economic Review* **93**(3):465–508.

Stern, N. 2007. *The Economics of Climate Change: The Stern Review*. New York: Cambridge University Press.

Wilcox, N. T. 2008. Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison. In *Risk Aversion in Experiments*, eds. J. C. Cox and G. W. Harrison. Greenwich, CT: JAI Press.

Wilcox, N. T. 2010. 'Stochastically More Risk Averse:' A Contextual Theory of Stochastic Discrete Choice Under Risk. *Journal of Econometrics*.

Williams, R. L. 2000. A Note on Robust Variance Estimation for Cluster-Correlated Data. *Biometrics* **56**:645–646.

Wooldridge, J. 2003. Cluster-Sample Methods in Applied Econometrics. *American Economic Review (Papers & Proceedings)* **93**:133–138.

CHAPTER 16

LABORATORY EXPERIMENTS: THE LAB IN RELATIONSHIP TO FIELD EXPERIMENTS, FIELD DATA, AND ECONOMIC THEORY

JOHN H. KAGEL

Introduction

The stated goal of this book, and the series it is part of, is to "formally open a dialogue on methodology," in this case for modern experimental economics. The particular topic I was asked to write about is lab experiments within the subheading "The Lab and the Field." In what follows I discuss laboratory experiments in relationship to field experiments and, more generally, in relationship to empirical research in economics and economic theory. All three empirical techniques—lab experiments, field experiments, and field data—are, in principle, fully complementary to each other and are informed by, and in turn inform, economic theory.

Furthermore, evaluating the results of laboratory experiments in relationship to results from other lab experiments, field experiments, field data, and theory are all critical in determining the external validity of lab results. My discussion will take place within the context of two case studies I am reasonably familiar with, one dealing with auctions, the other dealing with gift exchange in labor markets.

In writing about these relationships, it is important to keep in mind two things in terms of the context, or background, I bring to the discussion. First, behavior in experiments is often marked by a considerable learning period, which is no doubt true in field settings as well. In this context, one of the most important insights I have gained over the years is that whatever learning there is tends to be context dependent and difficult to generalize to new situations that are moderately, no less very, different from the learning context itself. This is not only my impression from the experimental results reported on here (which is rather overwhelming in the auction data I am most familiar with), but from reading the psychology literature explicitly designed to study learning and learning generalizability.[1] A second important insight is that oftentimes the same data say different things to different readers. That is, there is the experiment and then there is the interpretation of the results of the experiment—what the results mean in terms of the question(s) that motivated the study. It is not necessary to "buy" the author's interpretation of the results of an experiment in order to "buy" the experimental results. In fact, it is often conflicting interpretations of experimental results that lead to new experiments and better understanding of the phenomena in question, either by the initial experimenters or by competing groups. It is part of the healthy dialogue that is inherent to the research process.

The Winner's Curse

An Initial Set of Results and How They Relate to Theory and Experimental Methodology

The winner's curse story begins with Capen et al. (1971), three petroleum engineers who claimed that oil companies suffered unexpectedly low returns "year after year" in early Outer Continental Shelf (OCS) oil lease auctions. OCS auctions are common-value auctions, where the value of the oil in the ground is essentially the same to all bidders. Each bidder has their own estimate of the (unknown) value at the time that they bid. Even if these estimates are unbiased, bidders must account for the informational content inherent in winning the auction: The winner's estimate of value is (one of) the highest estimates. If bidders ignore this *adverse selection effect* inherent in winning the auction, it will result in below normal or even negative profits. The systematic failure to account for this adverse selection effect is referred to as the winner's curse: you win, you lose money, and you curse.[2]

Similar claims regarding a winner's curse have been made in a variety of other contexts: book publication rights (Dessauer, 1981), professional baseball's free agency market (Cassing and Douglas, 1980; Blecherman and Camerer, 1998), corporate takeover battles (Roll, 1986), and real estate auctions (Ashenfelter and Genesove, 1992). These claims have traditionally been greeted with a good deal of skepticism by economists as they imply that bidders repeatedly err, violating basic notions of rationality (which are unsustainable in the longer run). It is exceedingly difficult to support claims of a winner's curse with field data because of data reliability problems and plausible alternative explanations.[3]

The ambiguity inherent in interpreting field data, along with the controversial nature of the winner's curse, provided the motivation for laboratory studies of the winner's curse.[4] Bazerman and Samuelson (1983) conducted the first laboratory experiment demonstrating a winner's curse. Using M.B.A. students at Boston University, the experiment was conducted in class, with students participating in a series of first-price sealed-bid auctions in which bidders formed their own estimates of the value of each of four commodities: jars containing 800 pennies, 160 nickels, 200 large paper clips each worth four cents, and 400 small paper clips each worth two cents. Unknown to subjects, each jar had a value of $8.00. (Subjects bid on the value of the commodity, not the commodity itself.) In addition to their bids, subjects provided their best estimate of the value of the commodities and a 90% confidence bound around these estimates. A prize of $2.00 was given for the closest estimate to the true value in each auction. Auction group sizes varied between 4 and 26.

The average value estimate across all four commodities was $5.13 ($2.87 below the true value). In contrast, the average winning bid was $10.01, resulting in an average loss of $2.01 to the winner, with losses occurring in over half of all the auctions.[5] Further analysis showed that winning bidders were substantially more aggressive than other bidders, so that average winning bids were sensitive to a handful of grossly inflated bids.

The results of this experiment show that the winner's curse is easy to observe. However, many economists would object to the fact that subjects had no prior experience with the problem and no feedback regarding the outcomes of their decisions between auctions, so that the results could be attributed to the mistakes of totally inexperienced bidders. The robustness of these results is even more suspect given their sensitivity to a handful of grossly inflated bids, which one might suppose would be eliminated as a result of bankruptcies or learning in response to losses incurred in earlier auctions. In fact, it was just these objections that motivated my initial common value auctions experiments along with my colleague Dan Levin (Kagel and Levin, 1986; Kagel et al., 1989).

Our initial experiments were designed to correct for these problems: First, we conducted a *series* of auctions with cash payouts, preceded by two or three dry runs intended to deal with the experience issue. Second, we provided bidders with unbiased estimates of the value of the (fictitious) commodity they were bidding on, so

as to ensure that the affiliated values assumption underlying bidding in common value auctions would be satisfied and that winning bids yielding negative profits could not be attributed to biased estimates of the (unconditional) expected value of the item (as distinct from a biased estimate conditional on the event of winning). Third, following each auction we provided bidders with the full set of bids along with the corresponding signal values and the true value of the item, as well as the winning bidders profits, designed to speed up the learning process and to help subjects recognize the adverse selection problem (i.e., that winning bidders tend to have the highest, or one of the highest, estimates of the value of the item).

Our working hypothesis was that after a few rounds, bidders would settle down to earning a substantial share of the expected risk-neutral Nash equilibrium (RNNE) profits in these auctions. Boy, were we wrong. Auctions with inexperienced bidders showed a pervasive winner's curse (Kagel et al., 1989): Profits in the first nine auctions averaged –$2.57 compared to the RNNE prediction of $1.90, with only 17% of all auctions having positive profits. Note that this is after bidders had participated in 2–3 dry runs, with full feedback, so that the results cannot be attributed to a total lack of experience. These negative profits are not a simple matter of bad luck either, or a handful of grossly inflated bids, as 59% of all bids and 82% of the high bids were above the expected value conditional on winning the item. Furthermore, 40% of all subjects starting these auctions went bankrupt. In short, the winner's curse was alive and well even after our best efforts to establish an environment that would correct for the objections of most mainstream economists. These results have since been replicated by a number of others under similar conditions (e.g., Lind and Plott, 1991; Cox et al., 2001).[6]

Not to be deterred, we brought back experienced subjects who had participated in this initial experiment and began to look at some of the comparative static predictions of the theory (Kagel and Levin, 1986).[7] This was motivated by two thoughts: First, maybe subjects just needed more experience. Second, perhaps subjects were just miscalibrated, so that the comparative static predictions of the theory would be satisfied in spite of the overbidding, which would constitute a far less damaging mark against the underlying theory (in our minds at least) than simply overbidding. The results looked pretty good, at least for small numbers of bidders, as auctions with 3 or 4 bidders earned substantial positive profits averaging $4.32 per auction (significantly greater than zero, but still well below the RNNE prediction of $7.48 per auction). However, for these same bidders, bidding in larger groups (auctions with 6–7 bidders) profits averaged –$0.54 per auction, compared to the RNNE prediction of $4.82. Note that predicted profits decreased by $2.66 per auction while actual profits decreased by almost twice as much ($4.86), along with a sizable frequency of the winner's curse as measured by the high bid exceeding the expected value conditional on winning.

At the same time, we observed significant violations of two key comparative static predictions of the underlying theory. First, regressions showed that, other things being equal, individual bids increased in response to increased numbers of

rivals, a characteristic considered to be a telltale sign of a winner's curse in field data for first-price sealed-bid auctions. Second, public information in the form of announcing the lowest private estimate of the value of the item, which in equilibrium should increase sellers' profits (Milgrom and Weber, 1982), did so in auctions with 3–4 bidders. But this same public information had just the opposite effect in auctions with 6–7 bidders, reducing sellers' profits.

We reached several conclusions based on these results. First, the overbidding identified as a winner's curse was not just a matter of subjects being miscalibrated, but extended to violating key comparative static predictions of the theory, including a key public policy prescription that public information, if available, should be used to raise revenue in the sale of government assets with a significant common-value element. Second, the reemergence of a winner's curse in going from experienced bidders in auctions with 3–4 bidders to one with 6–7 bidders with its heightened adverse selection effect was the first indication that the underlying learning processes resulting in elimination of the worst effects of the winner's curse was context specific rather than involving some sort of "theory absorption" that readily generalized to new environments. There have since been other manifestations of this, most notably in almost common value auctions (Rose and Kagel, 2008).

Third, we showed a striking similarity between our experimental results and an important anomaly in the field data on common value auctions, thereby addressing the ever-present question in laboratory experiments of external validity: Rates of return calculated for drainage leases in OCS sales showed that *both* neighbors and non-neighbors earned higher rates of return on drainage compared to wildcat leases (Mead et al., 1983, 1984).[8] A wildcat lease is one for which no drilling data are available, so that bidders have symmetric information based on seismic readings. A drainage lease is one in which hydrocarbons have been located on an adjacent tract so that there is asymmetric information, with companies who lease the adjacent tracts (neighbors) having superior information to other companies (non-neighbors). Theory predicts that in equilibrium, with an asymmetric information structure of this sort, neighbors will earn substantially more, on average, than on wildcat leases, and non-neighbors will earn substantially less (essentially zero profits according to the standard model of common value auctions with "insider" information at the time).[9] Obviously, the field data do not square with this theoretical result. However, it is consistent with the laboratory results once one recognizes that there is a considerable amount of public information about the oil in the ground associated with drainage tracts.[10] Similar to the laboratory results, this public information may have corrected for a winner's curse that depressed rates of return on wildcat tracts. Now clearly, this is not the only possible explanation for the field data: The leading alternative explanation is that the lower rate of return on wildcat leases reflects the option value of the proprietary information that bidders planned to realize on neighbor tracts if hydrocarbons were found. However, what the explanation based on public information increasing bidders'

profits in the presence of a winner's curse has going for it is its consistency with the original claims of Capen et al. (1971). Furthermore, it's a much more straightforward explanation than that bidders planned to lose money on wildcat leases, only to make up for it some years later on drainage leases should hydrocarbons be found.

Further Questions Related to External Validity—Field Experiments of a Sort

Conventional laboratory experiments in economics typically use financially motivated students as subjects. One ongoing issue with this subject population of convenience is that perhaps this ignores important selection effects that jeopardize the external validity of the results reported. There are several dimensions to this argument: Students might be self-selected in some way so as to exclude individuals with characteristics that are important determinants of the behavior in the underlying subject population of interest. This could consist of any number of things including inexperience with the task at hand, general immaturity, and the failure to select out individuals with the relevant "smarts" to thrive in markets where poorly performing individuals are eliminated from the market as part of the competitive process.[11]

One possible solution to this issue is to compare the behavior of professionals who have experience with the institution of interest to students. Dyer et al. (1989) report one such experiment comparing bids of seasoned executives in the commercial construction industry to that of students. To enhance the ecological validity of the experiment, we employed a first-price low-bid-wins auction, the same as the competitive bidding environment the executives operated in. Other than that, the experimental procedures were the same as in the demand side auctions already discussed.[12]

Similar results were found between the construction executives and the students: Both suffered from a winner's curse by earning negative average profits, with winning bidders consistently bidding below the expected value conditional on winning. Furthermore, announcing the highest private information signal for the executives raised average offer prices by a statistically significant amount, whereas in equilibrium it should have lowered them. (This corresponds to the finding that public information lowered bids in high-bid auctions when bidders suffer from a winner's curse.) Regression results showed that the executives bid slightly higher in auctions with larger numbers of competitors, qualitatively in the direction predicted by the RNNE, but the magnitude of the response was not large enough to avoid even larger average losses with more bidders. This response to changing numbers of bidders is qualitatively different from the students in both low- and high-price auctions and may well represent a difference between the two subject pools as a result of the executives' past experience. However, there is an alternative explanation based on the losses the executives experienced in the small

numbers market (as opposed to the positive profits the *experienced* students had in the small markets). As such, with increased numbers of bidders, simple survival pressures would have required bidding less aggressively for the executives to avoid bankruptcy and exit from the market.

One can, of course, dismiss the executives suffering from a winner's curse on the grounds that they failed to take the experimental incentives seriously. But the data suggest otherwise as they were carefully attentive throughout and estimates of bid functions were qualitatively similar to the students as bids were monotonically increasing in signal values.

Further investigation (Dyer and Kagel, 1996) begins to resolve the apparent discrepancy between the executives suffering a winner's curse in the lab with their apparent success in field settings. Two possibilities, which are not necessarily mutually exclusive, were identified. One is that the executives had learned a set of situation-specific rules of thumb which enabled them to avoid the winner's curse in the field, but which could not be applied in the laboratory. For example, an important determinant of the risk associated with bidding a job, which impacts the cost estimates assigned to job components and the markup assigned to the contract, involves (a) the architect/owner's reputation for admitting mistakes, or ambiguities, in construction plans and (b) their willingness to implement functionally equivalent construction outcomes using alternative, and cheaper, construction techniques than originally specified.[13] Needless to say, the contractors did not have any of these situation-specific rules to rely on when bidding within the relatively abstract context of the laboratory experiment. This factor is fully consistent with the context-specific nature of learning that psychologists have identified and that we have seen in our auction experiments.

The second factor at play is that the bidding environment created in the experiment, which is based on theoretical work, is not fully representative of the environment encountered in the construction industry. These involve two factors: (i) repeated play elements which at times permit low bidders to withdraw their bids *without penalty* following a "reasonable" mistake in calculating the bid estimate and (ii) differences in bidder estimates of construction costs are a much smaller factor in the construction industry compared to OCS auctions, with overhead and the amount of idle resources anticipated often playing the decisive role in determining the low bid, so that the executives were not prepared for the large bid factor needed to avoid the winner's curse in the lab. As for this second factor, looking at the results of other experiments comparing students with professionals (for example, Burns (1985), Garratt et al., (2012), reviewed in Fréchette, (2015), this volume), I am pretty well convinced that the major insight gained from such studies is the ability to identify unanticipated differences between field settings and how we model these situations.

More recently, Harrison and List (2004, 2008) present results which, on some dimensions at least, appear to show how practicing professionals are able to avoid the winners curse in an "artifactual field experiment" as they have "… *developed a*

heuristic that "travels" from problem domain to problem domain."[14] In what follows I offer an alternative explanation for this result that is more in line with my argument regarding context-specific learning. Harrison and List also report results from a second experiment in which the same professionals dealing with a commodity they are familiar with, sports cards, provides "… *tentative support for the hypothesis that naturally occurring markets are efficient because certain traders use heuristics to avoid the inferential error that underlies the winner's curse*." With regard to this second experiment, I think that there is some misunderstanding of the nature of the winner's curse. More importantly, I think that there is some fundamental misunderstanding of how we view the results of our experiments identifying a winner's curse as they relate to behavior in "mature" markets.

In one experiment, Harrison and List (2008) compare bids of sports-card dealers with nondealers using the same abstract, induced valuation procedures reported on above. They employ both a symmetric information procedure in which all bidders have the same level of information about the common value (a single draw from a uniform distribution whose midpoint is the common value) as well as an asymmetric information treatment where one bidder, the insider, knows the true value with certainty while all other bidders draw an information signal from a symmetric distribution around the true value. In all cases, subjects bid in a single auction after having participated in a minimum of 10 practice auctions with either 4 or 7 bidders.[15] Dealers rarely suffer from a winner's curse in the symmetric information treatment as they typically bid below the expected value conditional on winning. In contrast, nondealers suffered from the typical winner's curse, bidding above the expected value conditional on winning and earning negative average profits. In the asymmetric information treatment, dealers (as "outsiders") suffered from a winner's curse for 24–30% of all observations. And they were unable to reject a null hypothesis of no difference in the bid functions for dealers and nondealers in their role as "outsiders." In short, dealers responded to the heightened adverse selection effect of an insider being present by bidding more aggressively than in the symmetric information case.

Harrison and List argue that the superior performance of dealers with symmetric information reflects field experience in comparable settings and that, unlike the construction contractors, "… context-specific experience does appear to carry over to comparable settings, at least with these types of auctions." They attribute the substantial increase in the frequency of the winner's curse for dealers in the asymmetric information treatment to the fact that "*when dealers are placed in an unfamiliar role they perform relatively poorly*." (Harrison and List, 2008, p. 835, italics in the original) So at best, their results provide an argument for very limited learning generalizability: Dealers having adapted to adverse selection effects in field settings with symmetric information do not recognize the heightened adverse selection effect when an insider is present, succumbing to the winner's curse.

While this result is consistent with my argument for context-specific learning, I believe that there is an alternative explanation for the symmetric information

results. Dealers in buying trading cards must purchase them at low enough prices to be able to sell them at a profit and would most certainly be in the habit of doing so; for example, List and Lucking-Reiley (2000) show that dealers bid just under $50 for cards with a retail value of $70 in a Vickrey auction. So applying such large discount factors, which nondealers in their typical role as buyers would not be in the habit of doing, could very well protect them from a winner's curse.[16] There is independent support for this interpretation of their results from an experiment by Garrat et al. (2004), who report that subjects with strong experience as sellers in eBay auctions systematically bid below their *induced* values in a single-unit second-price sealed-bid auction.[17] On this interpretation, the "*heuristic that "travels" from problem domain to problem domain*" is that dealers have learned to buy low and sell high which adventitiously, in this case, protects them from the winner's curse.[18]

In Harrison and List's (2008) second experiment, the professional card dealers are bidding on a commodity they are familiar with, an *unopened* package of *Leaf* sports cards (packages containing 10 cards of unknown value, and an *established* retail price of between $9 and $10). In comparing bids of dealers with nondealers, there is not a single bid above $10 for dealers and only a handful of bids above $10 for nondealers, so that not a single dealer bid above the established retail price, but a few nondealers did.[19] Average winning bids were around $5 for dealers and a little over $8 for nondealers. These dealer bids (never exceeding $10) appear to underlie the claim "… that naturally occurring markets are efficient because certain traders use heuristics to avoid the inferential error that underlies the winner's curse." While we agree the Leaf sports card auctions constitute a common value auction, it is one in which there is no scope for a winner's curse. The cards have a well-known market value, which precludes any adverse selection effect based on different estimates of their value, particularly on the part of professional card dealers. As such, for the dealers at least, this experiment is comparable to auctioning off a $9 or $10 bill, which they in turn expect to resell.

But there is a broader claim here regarding the efficiency of bidding in established markets and the absence of a winner's curse that requires some clarification. Finding a winner's curse with inexperienced bidders in the laboratory, in conjunction with field reports of a winner's curse (e.g., Capen et al., 1971), along with the parallels between field data and laboratory data reported on above, leave us reasonably well convinced that a winner's curse was present in early OCS auctions and is likely to exist *at least in the start up* phase of auction markets with a strong common-value element. Furthermore, if our interviews with the construction contractors are at all representative, it's very likely to be present for new entrants into those markets, or even experienced agents when entering a new segment of the market (Dyer and Kagel, 1996). But as the laboratory experiments have also shown, people do learn and there are market selection effects so that one would *not* expect long established players in any given market to suffer from a winner's curse on a *regular* basis.

Gift Exchange in Experimental Labor Markets

Laboratory auction experiments are generally considered to be a reasonably good model for their field counterparts in that underlying structure of the lab and target environments are similar; for example, a first-price sealed-bid auction in the lab has the same rules as its field counterpart. That is, the empirical interpretation of the concepts and models underlying auctions are quite similar between laboratory and field settings. To be sure, there are still important differences. For example, laboratory investigations of the winner's curse typically employ a single-unit pure common-value auction environment with a known number of bidders, whereas field settings often involve a mix of common- and private-value elements, with an unknown number of actual bidders, frequently demanding multiple units. But to the extent that there is a strong common value element in the field setting with a number of competing bidders, there is a strong adverse selection effect that bidders' must contend with in both cases. Thus, the insights about behavior gained in the laboratory can, and have, provided insights into field settings: One of the arguments in favor of open, ascending price auctions for the Federal Communication Commission's sale of spectrum (air-wave) rights was that "… allowing bidders to respond to each other's bids diminishes the winner's curse: that is, the tendency for naive bidders to bid up the price beyond the licenses' actual value, or for shrewd bidders to bid cautiously to avoid over paying" McAfee and McMillan (1996, p. 161). Furthermore, to the extent that laboratory experiments are used as a testbed for designing auctions to be implemented in field settings (e.g., Plott, 1997; Ledyard et al., 1997), there are special reasons to think that the experiment provides a good working model for the target environment and, hence, for the external validity of the results reported.

The Relevance of Gift Exchange Experiments for Understanding Labor Relations in Firms: Background Material

There is considerably greater controversy regarding the correspondence between the empirical concepts underlying laboratory studies of gift exchange in labor markets relative to the target environment of labor relations in firms (see Bardsley et al., 2009, for example). However, before addressing this issue, I provide some background concerning motivation for these experiments and summarize some of the basic laboratory results and related field experiments, along with some of the methodological issues and interpretation issues that arise in these comparisons.

Akerlof (1982) was the first to introduce the idea of gift exchange in labor markets; firms paying above market wages because of the negative effects of low wages

on worker morale, resulting in less shirking as a reciprocal response on the part of workers. Among other things, the model has been used to explain the reluctance of firms to cut nominal wages as unemployment increases in an economic downturn. Of course, there are many competing models designed to explain this last phenomena (see Bewley (2005), for a review).

The idea of gift exchange as a reciprocal response of employees to higher wages is one that has been extensively studied, originally in a series of laboratory experiments pioneered by Fehr and his colleagues (e.g., Fehr et al., 1993). The typical gift exchange game is a two-stage game. In stage 1, employers' make costly wage offers to potential employees. In stage 2, employees decide to accept or reject the proposed offer and then provide a costly "effort level" to employers, with more effort being more costly. The higher the effort level provided, the greater the employer's profits. In most early lab experiments, matching of firms and workers was anonymous, taking place over a finite number of trials announced in advance, with different partners in each trial, all taking place well within two hours, so that there is no opportunity for workers to develop individual reputations or for other repeated play game effects. "Effort" is defined in terms of "employees" choosing higher numbers which reduce their experimental earnings while increasing the earnings of their "employer."

Since there are a finite number of plays of the game with a known end point, standard unraveling arguments predict wage offers at the bottom end of the range permitted in all periods, accompanied by minimum effort levels. Yet there is typically a clear positive relationship between wages and effort levels resulting in a Pareto improving outcome with higher earnings for both firms and workers than under the competitive equilibrium outcome. These results do not appear to depend on the fine points of the market institutions (e.g., posted offer versus double oral auction or one-to-one matching) and are robust to introducing a real effort task into environment (e.g., Gneezy (2004), who used solving mazes as his real effort task).

The role of positive and negative reciprocity between workers and firms in this setting is reasonably noncontroversial. For example, Fehr and Falk (1999) look at gift exchange in very demanding double auction markets. In one treatment, with opportunities for gift exchange which could, potentially, improve outcomes for both "employers" and "employees," employers offer "wages" above the competitive equilibrium level and spurn efforts of employees to undercut these wages, as they anticipate lower effort levels. However, with exogenously determined effort levels, employers accept only the lowest wage offers, with wages forced down to close to the competitive equilibrium level.[20] Their preferred interpretation of these results is that employers' behavior is driven mainly by the expectation of positive reciprocity on the part of workers to higher wages, as opposed a simple desire to raise workers earnings out of "fairness" considerations.[21] To take another example, Charness (2004) compares effort levels in a typical two-stage gift exchange experiment with one-to-one matching of employers with employees. In one treatment, "wages" are determined exogenously, either randomly or by the experimenter, with

the distribution of these exogenous wage offers designed to match the distribution observed with employer determined wages. There are higher effort levels with exogenously determined as opposed to employer determined wage offers at lower wage rates, with no difference at higher wage rates. He interprets this as evidence for negative reciprocity on the part of workers as they respond with the minimum possible effort to low employer determined wages. He interprets the absence of higher effort levels in response to higher employer determined wages as evidence against positive reciprocity. However, this is not the only possible interpretation, with other investigators reporting evidence for positive reciprocity at higher wage rates (Owens and Kagel, 2009).[22]

To be sure, there are variations in effort levels reported in different experiments. For example, Hannan et al. (2002) report marked differences in reciprocity between undergraduate students and MBAs under the same experimental conditions, with MBAs providing significantly higher effort levels at comparable wage rates, as well as providing higher average wages. They conjecture that the difference between MBAs and undergraduates results from MBAs greater experience in jobs where gift exchange plays an important role, thereby sensitizing the MBAs to the reciprocity considerations inherent in higher wage offers. In contrast, most undergraduate work in the United States is associated with minimum wage jobs where there is no, or minimal, gift exchange. There is some support for this conjecture in a subsequent study with the same subject population, as MBAs with above-median professional work experience provided a higher mean effort compared to subjects with below-median professional work experience (Hannan, 2005). Hannan et al. (2002) also report no significant difference in effort levels of workers in response to comparable wage offers from high versus low productivity firms (where high productivity firms find it less costly to provide higher wages than do low productivity firms).[23] They conjecture that this lack of responsiveness, which was true for both MBAs and undergraduates, is due to the fact that the relationship between firm profits and productivity is an indirect one, hence not as likely to be as salient to subjects as differences in wage offers. In a subsequent experiment, Charness et al. (2004) found that the degree of gift exchange is surprisingly sensitive to how payoffs are framed: They find substantially less effort provided (among undergraduates) when a comprehensive payoff table relating wages and effort levels to workers' payoffs and employers' earnings was provided as part of the instructions as opposed to the usual format of providing payoff functions and a set of examples for subjects to work through.[24] However, similar payoff tables had no deleterious effect on MBAs effort levels compared to results typically reported (Hannan et al., 2002).

In a field experiment, Gneezy and List (2006) question the staying power of positive reciprocity in response to higher than anticipated wage rates. They looked at two tasks: computerizing library holdings over a 6-hour period and a door-to-door fundraising effort over a single weekend day. The gift exchange treatment was operationalized by advertising a given wage rate, with higher than advertised

wages paid to one of two treatment groups—for example, an advertised wage of $12 per hour for the library task, with half the subjects given the "surprise" wage of $20 an hour upon showing up. They break their data in half, reporting significantly higher effort levels for the high-wage group in the first half of the day, but no significant differences in the last half of the day. They interpret these results in terms of the psychology literature on reference point effects, arguing that after a while a worker's reference point shifts so that the new higher wage serves as the fair-wage reference point, thereby generating lower effort. Clearly, this is not the only possible interpretation of their results: Perhaps the higher-wage workers became fatigued from working harder and/or the higher-wage workers provided the gift level they thought appropriate to the higher than advertised wage in the first half of the day and slacked off after that.[25]

Subsequent field experiments report positive gift exchange between firms and workers that in a number of cases is *increasing* over time. Kube et al. (2013) look at gift exchange in a library cataloguing task, focusing on both positive and negative reciprocity. Students were hired for a six-hour shift to catalogue books, with the recruitment e-mail announcing a *presumptive* salary of 15 euros per hour. Upon arrival, one-third of the subjects were told that the wage would be 20 euros per hour (the "Kind" treatment), with another one-third told that the wage was actually 10 euros per hour (the "Unkind" treatment) and the last group would receive the 15 euro per hour wage (the "Neutral" treatment). The number of books catalogued increased over time for all three treatments. The Unkind treatment starts out at a much lower rate of cataloguing than the Neutral treatment and remains below it throughout, with the differences statistically significant in each 90-minute interval. The Kind treatment starts at the same rate as the Neutral treatment, but it catalogues at higher rates for each 15-minute interval after that. However, with the exception of the middle interval, these differences are not statistically significant at conventional levels. Kube et al. (2006) conclude that negative reciprocity is a stronger force than positive reciprocity, which may well be the case. But the small sample sizes involved (9 and 10 subjects in each treatment), in conjunction with the time pattern observed in the data between the Kind and Unkind treatments, would seem to be a shaky basis to reject the existence of positive reciprocity in field experiments of this sort. Indeed, Al-Ubaydli et al. (2007) provide evidence in favor of positive reciprocity in a field experiment with temporary workers in which effort levels are growing over time in both the baseline treatment and the positive gift exchange treatment, with the differences in favor of the gift exchange treatment increasing over time, to the point that they appear to be statistically significant at the end of their two day trial.[26] Finally note that both experiments report the opposite time pattern between the controls and the gift exchange treatment when compared to the one that Gneezy and List reported, which is contrary to their conjecture of the behavioral process underlying their results.

The discussion so far illustrates both the strengths and weaknesses of laboratory and field experiments on this topic. Subjects in field experiments do not know

they are in an experiment, the tasks they are being asked to perform parallel those they would normally perform, and, in some cases, the experiment takes place in the context of a longer-term relationship.[27] As such, these experiments should have a closer relationship to the field setting that authors such as Akerlof had in mind. However, the cost of this verisimilitude is high. There is considerable loss of control in these experiments, as we neither know the cost of effort, the perceived benefits of effort to the employer, nor the game that the employees think they are playing. Measurement is a problem in many of these studies as workers in field settings can respond to incentives along multiple dimensions, so that the experimenter may miss important elements of employees' responses to higher wage rates. Also one must account for the level of baseline wages relative to market wages for comparable work, as higher-than-normal baseline wages may already elicit a strong gift response. As Cohn et al. (in press) note, this may impose a ceiling effect resulting in a downward bias in the response to the gift wage treatment, and baseline wages in relationship to market wages are not always reported.

So What Do Any of These Experiments Have to Tell Us About the Gift Exchange and Sticky Downward Wages in the Workplace?

None of the experiments reported on so far directly address the question of sticky downward wages to clear unemployment in a recession.[28] Furthermore, the very structure of these experiments, both laboratory and field studies, cannot approach the target environment of ongoing labor relations within firms, no less one in which there is a downturn in economic activity for the economy. In addition, at least one prominent student of survey research on why firms typically do not cut wages during recessions (Bewley, 2005) seriously doubts the empirical relevance of the morale theory of wage rigidity proposed by Akerlof (1982) and others (Solow, 1979; Akerlof and Yellen, 1988, 1990). Bewley does so on the grounds that "…. employers say that they do not see much connection between effort or morale and wage levels; effort and morale do not increase with pay levels …" (Bewley, 2005, p. 309). He goes on to note that what is accurate in the theory is that employers avoid cutting wages because to do so would hurt morale. In contrast, higher wages per se have little effect because employees usually have little notion of a fair market value for their services, but quickly come to believe that they are entitled to their existing pay, no matter how high it may be.

Nevertheless, Bewley sees considerable relevance for the gift exchange experiments described here to the issue at hand—sticky downward wages. He does so on the grounds that in lab experiments:

> The most important finding is the prevalence of reciprocity. …. The general willingness to reciprocate good to good is the essence of good morale. Negative reciprocity is what underlies the insult effect of pay cuts, which is resentment

> caused by the firm's perceived breach of positive reciprocity; workers expect employers to offer pay increases, not cuts, in exchange for loyalty and effort. (Bewley, 2005, p. 318)

Thus, although there is a certain unreality associated with the structure of both laboratory and field experiments dealing with gift exchange in labor markets in relationship to the field settings they are designed to provide insight into, what the experiments isolate and validate is the prevalence of reciprocity, the core principle underlying sticky downward wages. It is on the basis of this, rather than a fully faithful empirical model of the target environment (an impossibility under any circumstance), that makes for the relevance of the experiments for the target environment.

Summary and Conclusions

I have discussed the use of laboratory experiments and their relationship to field data, field experiments, and economic theory. I have done this in the context of two examples I am reasonably familiar with: (a) common-value auctions and the winner's curse and (b) gift exchange in experimental labor markets. I have tried to make three overarching substantive points.

First, learning, which is endemic to most experimental studies, tends to be context specific and difficult to generalize from one environment to another. This is totally consistent with the psychological literature on learning which distinguishes between near transfer (e.g., someone who knows how to drive a car can typically handle driving a light truck) and far transfer (e.g., knowing how to drive a car does not transfer immediately to driving a semi-trailer, no less an airplane or speedboat). This has a number of implications. Among other things, one should not automatically expect economic agents who are fine tuned to the field settings they typically operate in to be able to adjust to experimental environments that deprive them of their typical contextual cues or that vary even in small, but important, ways from their natural habitat. And although agents in an experiment have converged on an equilibrium in a given economic setting, this does not necessarily imply that they will respond in the way that theory suggests, without a new round of learning, when the environment changes.

Second, although hopefully we can all agree on the "facts" of a particular economic experiment, there is typically wide room for disagreement on the interpretation of those facts as they apply to the broader issues at hand. I've supplied two examples on this point, one for each of the two cases I've considered. (I could provide many more as well in any number of areas of active experimental research past and present.) I don't expect everyone to agree with my alternative interpretations, no less the investigators that reached the original conclusions. In fact it's

these differences in how one interprets the facts that leads to new and interesting experiments that further narrow what we do and don't know on a particular topic.

Third, I have tried to show that even in cases where the laboratory setting seems rather removed and abstract relative to the field setting one has in mind, the experimental results may be quite relevant to that field setting. The trick here is that the experiment, as abstract and as seemingly far removed from the target environment in question, needs to capture the essential behavioral issue(s) at hand: Do real agents behave in the ways our theories predict and for the reasons the theory postulates? Coming back to my second point, there will no doubt be some who will disagree with my interpretation on this point for the case in question.

NOTES

Special thanks to my long-time co-authors Dan Levin and David Cooper because the ideas expressed here have been largely developed in working with them: special thanks also to participants in the New York University Conference on Methods of Modern Experimental Economics, especially my discussant David Reiley. Research has been partially supported by National Science Foundation grants SES-0452911, 0851674, and 0924764. Any opinions, findings, conclusions, or recommendations in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. I alone am responsible for errors and omissions.

1. There is a whole body of psychological literature indicating the difficulty of learning generalizing across different contexts (see, for example, Gick and Holyoak (1980), Perkins and Salomon (1988), and Salomon and Perkins (1989)).

2. Unfortunately, many economists, particularly theorists, characterize the winner's curse as the difference between the expected value of the item conditional on winning and the unconditional, naive expectation, using the term to refer to bidders fully accounting for this difference, rather than failing to do so and losing money as a consequence. This can make for some confusion.

3. For example, Hendricks et al. (1987) found that in early OCS lease sales, average profits were negative in auctions with seven or more bidders. Hendricks et al. note that one possible explanation for this outcome is the increased severity of the adverse selection problem associated with more bidders. However, they note that the data could also be explained by bidder uncertainty regarding the number of firms competing on a given tract (their preferred explanation) so that bidders discount by the usual number of rivals, which is less than seven, earning negative profits as a result of the failure to accurately anticipate the number of rivals. This represents an adverse selection effect of a different sort from the winner's curse.

4. One question that comes up with field data is the extent to which a particular auction is dominated by private as opposed to common value elements. In the lab, one can ensure a pure common-value or private-value auction, or some interesting combination in between.

5. Winning bidders paid these losses out of their own pockets or from earnings in the other auctions (Max Bazerman, personal communication).

6. At the same time there were a parallel series of laboratory experiments demonstrating a robust winner's curse in the corporate takeover game first reported in

Samuelson and Bazerman (1985), results of which are reviewed in Kagel (1995) and Kagel and Levin (2008); also see Charness and Levin (2009).

7. The curious juxtaposition of the paper with experienced bidders being published before the paper with inexperienced bidders is explained by the vagaries of the publication process.

8. Hendricks and Porter (1992) obtained net rate of return estimates quite similar to those of Mead et al. (1983, 1984) on this score as well.

9. See Wilson (1967), Weverbergh (1979), Engelbrecht-Wiggans et al. (1983), and Hendricks et al. (1994) for analysis of the standard model.

10. See Cooper (1998) for discussion of the extensive spying that goes on between rival companies once drilling starts on a tract and the difficulties involved in keeping drilling results out of the hands of competitors.

11. This is quite distinct from the question of representativeness, as in cases where one wants to use lab experiments as a reliable way of measuring field preferences for a much more heterogeneous population. See, for example, Andersen et al. (2010) or Bellemare et al. (2007).

12. In terms of Harrison and List's (2004) proposed taxonomy of field experiments, this is an "artifactual field experiment," the same as a "conventional" laboratory experiment, employing abstract framing and an imposed set of rules but using a nonstandard subject pool, professionals who presumably have some experience with the relevant institution, thereby alleviating some of the potential selection effects associated with student subjects.

13. This is important enough to the point that in the experiment, at least one of the executives jokingly inquired, "Who is the architect associated with this job?"

14. See Harrison and List (2004, p. 1027) for both quotes (italics in the original).

15. Subjects were not provided with any starting capital balances or participation fees to cover potential losses. However, a second experiment that was not announced until after the first was completed was used to ensure that everyone earned positive profits.

16. In that experiment, List and Lucking-Reiley (2000) assume that the trading card market is best approximated by a private-value auction.

17. In contrast, buyers whose primary experience with eBay auctions involved buying will typically bid above their induced values!

18. One should, of course, ask why this same heuristic does not help the construction contractors. The answer is simple. General contractors do not buy and sell in anything approaching the same way that card dealers do. Rather, they solicit bids from large numbers of subcontractors who are responsible for fulfilling their commitments and then add in their own estimated general contactor's costs.

19. I am referring to what they call the SIS auctions—dealers bidding against other dealers and nondealers bidding against other nondealers.

20. There is zero worker surplus/profits at the competitive equilibrium so that wages are a bit higher than the competitive equilibrium itself.

21. An alternative explanation is that firms are concerned with negative reciprocity should they accept the low wage offers.

22. The issue of whether positive or negative reciprocity plays a stronger role in subject behavior is still under debate; see, for example, Offerman (2002) and Cox et al. (2008) for contrasting points of view. Hannan (2005) provides evidence that employees reduce effort levels more in response to a reduction in wage rates following a negative profit shock than the increase in effort in response to an increase in wage rates following a positive profit shock.

23. Wage offers were tagged with the firm's productivity level in a posted offer labor market.

24. Note that there was a positive relationship between effort levels and wages as typically reported in laboratory gift exchange experiments; also note that average wages and effort levels were significantly lower with the payoff table than without it.

25. A number of writers have cited Hennig-Schmidt et al. (2010) as support for the Gneezy–List results. What Hennig-Schmidt et al. report is an absence of positive reciprocity to higher than anticipated wages in a lab experiment with real work effort and a corresponding field experiment. They hypothesize that this negative outcome results from the absence of explicit cost and surplus information that would enable employees to calculate employer's surplus. An additional real-effort lab treatment supports this hypothesis.

26. Unfortunately, the statistical specification does not include any interaction term for the time trend variable with the treatment effects which would reveal whether output levels in the gift exchange treatment were growing significantly faster over time than the controls, nor is the data split into early versus late responses as in Gneezy and List (2006).

27. In some quarters at least, not reporting to subjects that they are part of a field experiment of the sort described here is unacceptable to human subject committees.

28. The closest experiment I have seen dealing with this issue is Hannan (2005).

REFERENCES

Akerlof, G. A. 1982. Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics* **97**:543–569.

Akerlof, G. A. and J. Yellen. 1988. Fairness and Unemployment. *American Economic Review: Papers and Proceedings* **78**:44–49.

Akerlof, G. A. and J. Yellen. 1990. The Fair Wage-Effort Hypothesis and Unemployment. *Quarterly Journal of Economics* **105**:255–283.

Al-Ubaydli, O., S. Andersen, U. Gneezy, and J. A. List. 2007. For Love or Money? Testing Non-pecuniary and Pecuniary Incentive Schemes in a Field Experiment. Working paper, University of Chicago.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutstrom. 2010. Preference Heterogeneity in Experiments: Comparing the Field and the Laboratory. *Journal of Economic Behavior and Organization*.

Ashenfelter, O. and D. Genesove. 1992. Testing for Price Anomalies in Real Estate Auctions. *American Economic Review: Papers and Proceedings* **82**:501–505.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden. 2009. *Experimental Economics: Rethinking the Rules*. Princeton, NJ: Princeton University Press.

Bazerman, M. H. and W. F. Samuelson. 1983. I Won the Auction But Don't Want the Prize. *Journal of Conflict Resolution* **27**:618–634.

Bellemare, C., S. Kroger, and A. van Soest. 2007. Preferences, Intentions, and Expectations: A Large-Scale Experiment with a Representative Subject Pool. IZA DP No. 3022.

Bewley, T. 2005. Fairness, Reciprocity, and Wage Rigidity. In *Moral Sentiments and Material Interests: the Foundations of Cooperation in Economic Life*, eds. H. Gintis, S. Bowles, R. T. Boyd, and E. Fehr. Cambridge, MA: MIT Press.

Blecherman, B. and C. F. Camerer. 1998. Is There a Winner's Curse in the Market for Baseball Players? Mimeograph, Brooklyn Polytechnic University.
Burns, P. 1985. Experience and Decision Making: A Comparison of Students and Businessmen in a Simulated Progressive Auction, In *Research in Experimental Economics*, Volume 3, ed. Vernon L. Smith. Greenwich: JAI Press.
Capen, E. C., R. V. Clapp, and W. M. Campbell. 1971. Competitive Bidding in High-Risk Situations. *Journal of Petroleum Technology* **23**:641–653.
Cassing, J. and R. W. Douglas. 1980. Implications of the Auction Mechanism in Baseball's Free Agent Draft. *Southern Economic Journal* **47**:110–121.
Charness, G. 2004. Attribution and Reciprocity in an Experimental Labor Market. *Journal of Labor Economics* **22**:665–688.
Charness, G., G. Fréchette, and J. H. Kagel. 2004. How Robust is Laboratory Gift Exchange? *Experimental Economics* **7**:189–205.
Charness, G. and D. Levin. 2009. The Origin of the Winner's Curse: A laboratory Study. *American Economic Journal: Microeconomics* **1**:207–236.
Cohn, A., E. Fehr, and L. Goette. Fair Wages and Effort Provision: Combining Evidence from the Lab and the Field. Forthcoming in *Management Science.*
Cooper, Christopher. 1998. Oil Firms Still Rely on Corporate Spies to be Well-Informed. *Wall Street Journal* **December 7**:1, 23.
Cox, J. C., S. Dinkin, and J. T. Swarthout. 2001. Endogenous Entry and Exit in Common Value Auctions. *Experimental Economics* **4**:163–181.
Cox, J. C., K. Sadiraj, and V. Sadiraj. 2008. Implications of Trust, Fear and Reciprocity for Modeling Economic Behavior. *Experimental Economics* **11**:1–24.
Dessauer, J. P. 1981. *Book Publishing*. New York: Bowker.
Dyer, D., J. H. Kagel, and D. Levin. 1989. A Comparison of Naive and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis. *Economic Journal* **99**:108–115.
Dyer, D. and J. H. Kagel. 1996. Bidding in Common Value Auctions: How the Commercial Construction Industry Corrects for the Winner's Curse. *Management Science* **42**:1463–1475.
Engelbrecht-Wiggans, R., P. R. Milgrom, and R. J. Weber. 1983. Competitive Bidding and Proprietary Information. *Journal of Mathematical Economics* **11**:161–169.
Fehr, E. and A. Falk. 1999. Wage Rigidity in a Competitive Incomplete Contract Market. *Journal of Political Economy* **107**:106–134.
Fehr, E., G. Kirchsteiger, and A. Riedl. 1993. Does Fairness Prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics* **108**:437–460.
Fréchette, G. R. 2015. Laboratory Experiments: Professionals Versus Students. In *Handbook of Experimental Economic Methodology,* eds. Frechette, G. R. and A. Schotter. Oxford University Press.
Garratt, R., M. Walker, and J. Wooders. 2012. Behavior in Second-Price Auctions by Highly Experienced eBay Buyers and Sellers. *Experimental Economics* **15**(1):44–57.
Gick, M. L. and K. J. Holyoak. 1980. Analogical Problem Solving. *Cognitive Psychology* **12**:306–355.
Gneezy, U. 2004. Do High Wages Lead to High Profits? An Experimental Study of Reciprocity Using Real Effort. Working paper, University of Chicago, Graduate School of Business.
Gneezy, U. and J. List. 2006. Putting Behavioral Economics to Work: Field Evidence of Gift Exchange. *Econometrica* **74**:1365–1384.

Hannan, R. L. 2005. The Combined Effect of Wages and Firm Profit on Employee Effort. *The Accounting Review* **80**:167–188.

Hannan, R. L., J. Kagel, and D. Moser. 2002. Partial Gift Exchange in Experimental Labor Markets: Impact of Subject Population Differences, Productivity Differences, and Effort Requests on Behavior. *Journal of Labor Economics* **20**:923–951.

Harrison, G. W. and J. A. List. 2004. Field Experiments. *Journal of Economic Literature* **42**:1013–1059.

Harrison, G. W. and J. A. List. 2008. Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner's Curse. *Economic Journal* **118**:822–843.

Hendricks, K. and R. H. Porter. 1992. Bidding Behavior in OCS Drainage Auctions: Theory and Evidence. Presentation, 1992 European Economics Association meetings.

Hendricks, K., R. H. Porter, and B. Boudreau. 1987. Information, Returns, and Bidding Behavior in OCS Auctions: 1954–1969. *The Journal of Industrial Economics* **35**:517–542.

Hendricks, K., R. H. Porter, and C. A. Wilson. 1994. Auctions for Oil and Gas Leases with an Informed Bidder and a Random Reservation Price. *Econometrica* **62**:1415–1444.

Hennig-Schmidt, H., B. Rockenbach, and A. Sadrieh. 2010. In Search of Workers Real Effort Reciprocity—a Field and Laboratory Experiment. *Journal of the European Economic Association*, **8**:817–837.

Kagel, J. H. 1995. Auctions: A Survey of Experimental Research. In *Handbook of Experimental Economics*, eds. A. E. Roth and J. H. Kagel. Princeton, NJ: Princeton University Press.

Kagel, J. H. and D. Levin. 1986. The Winner's Curse and Public Information in Common Value Auctions. *American Economic Review* **76**:894–920.

Kagel, J. H. and D. Levin. 2008. Auctions: A Survey of Experimental Research, 1995–2008. Working paper, Ohio State University.

Kagel, J. H., D. Levin, R. Battalio, and D. J. Meyer. 1989. First-Price Common Value Auctions: Bidder Behavior and the Winner's Curse. *Economic Inquiry* **27**:241–258.

Kube, S., M. A. Maréchal, and C. Puppe. 2013. Putting Reciprocity to Work—Positive Versus Negative Responses in the Field. *Journal of the European Economic Association* **11**:853–870.

Ledyard, J. O., D. P. Porter, and A. Rangel. 1997. Experiments Testing Multi-object Allocation Mechanisms. *Journal of Economics and Management Strategy* **6**:639–675.

Lind, B. and C. R. Plott. 1991. The Winner's Curse: Experiments with Buyers and with Sellers. *American Economic Review* **81**:335–346.

List, J. A. and D. Lucking-Reiley. 2000. Demand Reduction in Multi-unit Auctions: Evidence from a Sportscard Field Experiment. *American Economic Review* **90**:961–972.

McAfee, R. P. and J. McMillan. 1996. Analyzing the Airwaves Auction. *Journal of Economic Perspectives* **10**:159–176.

Mead, W. J., A. Moseidjord, and P. E. Sorensen. 1983. The Rate of Return Earned by Leases Under Cash Bonus Bidding in Ocs Oil and Gas Leases. *Energy Journal* **4**:37–52.

Mead, W. J., A. Moseidjord, and P. E. Sorensen. 1984. Competitive Bidding Under Asymmetrical Information: Behavior and Performance in Gulf of Mexico Drainage Lease Sales, 1954–1969. *Review of Economics and Statistics* **66**:505–508.

Milgrom P. R. and R. J. Weber. 1982. A Theory of Auctions and Competitive Bidding. *Econometrica* **50**:1089–1112.

Offerman, Theo. 2002. Hurting Hurts More than Helping Helps. *European Economic Review* **46**:1423–1437.

Owens, M. F. and J. H. Kagel. 2009. Minimum Wage Restrictions and Employee Effort in Labor Markets with Gift Exchange Present. Working paper, Ohio State University.
Perkins, D. N. and G. Salomon. 1988. Teaching for Transfer. *Educational Leadership* **46**:22–32.
Plott, C. 1997. Laboratory Experimental Test Beds: Application to the PCS Auction. *Journal of Economics and Management Strategy* **6**:605–638.
Roll, R. 1986. The Hubris Hypothesis of Corporate Takeovers. *Journal of Business* **59**:197–216.
Rose, S. L. and J. H. Kagel. 2008. Bidding in Almost Common Value Auctions: An Experiment. *Journal of Economic Management Strategy* **17**:1041–1058.
Salomon, G. and D. N. Perkins. 1989. Rocky Roads to Transfer: Rethinking Mechanisms of a Neglected Phenomenon. *Education Psychologist* **24**:113–142.
Samuelson, W.F. and M. H. Bazerman. 1985. The Winner's Curse in Bilateral Negotiations. In *Research in Experimental Economics*, Volume 3, ed. V. L. Smith. Greenwich, CT: JAI Press.
Solow, R. M. 1979. Another Possible Source of Wage Stickiness. *Journal of Macroeconomics* **1**:79–82.
Weverbergh, M. 1979. Competitive Bidding with Asymmetric Information Reanalyzed. *Management Science* **25**:291–294.
Wilson, B. R. 1967. Insider Competitive Bidding with Asymmetric Information. *Management Science* **13**:816–820.

CHAPTER 17

LABORATORY EXPERIMENTS: PROFESSIONALS VERSUS STUDENTS

GUILLAUME R. FRÉCHETTE

Introduction

Most economic theories make general claims. That is, economic models rarely start with assumptions such as: The firms are picnic table producers in Korea. . . . Hence, whenever one engages in applied work testing a model, there will be issues of generalizability of the results to other data sets, environments, subject population, and so on. Such concerns extend to experimental data sets since, in particular, subjects tend to be undergraduate students while much of economic activity outside the laboratory is done by professionals. This chapter establishes what is known about the impact of the specific subject sample typically used in experimental economics by reviewing prior studies that have used both the standard subject pool and an unusual pool of professionals. Hence, two questions immediately come to mind: (1) Why make comparisons across subject pools? and (2) What is the usual subject pool and what are professionals?

What are some of the reasons why a researcher might want to study a nonstandard subject pool?

1. Substantive questions: Some fact or phenomenon about a particular group is observed outside of the laboratory, and the lab is used to identify the

cause(s) of that phenomenon. An example of this is the work of Niederle and coauthors looking at gender and how it interacts with competition and other factors relevant for salaries and careers of females versus males (Gneezy et al., 2003).

2. Comparison of parameter estimates across subject pools. For example: Are professional investors more risk averse than the usual experimental subjects (Potamites and Zhang, 2012)? How does risk aversion differ in the standard subject pool and in a representative sample of the population (Andersen et al., 2009)?
3. Would you reach a similar conclusion about behavior in a game or about the predictions of a theory if you used the usual experimental subjects versus if you used subjects who are career professionals at the task the game being tested is supposed to represent? An example of this is the work of Kagel and coauthors looking at how professionals from the construction industry bid in common-value auctions (as opposed to the usual experimental subjects). In particular, do professionals also fall prey to the winner's curse (Dyer et al., 1989)?
4. Using animals to do experiments that cannot be done with human subjects. An example of this is the work of Battalio, Kagel, and coauthors using pigeons and rats (Kagel et al., 1995).
5. There are other cases that do not fit neatly in the above categories; here are two examples: (a) Comparing children to adults; see the work of Harbaugh and coauthors (Harbaugh et al., 2001). Difference in behavior across subgroups in an experiment where this is observed ex post but is not a purpose of the original study; for example, Casari et al. (2007) observed differences in bidding behavior between males and females in an auction experiment.

Readers interested in issues of subject pool differences in general are referred to Ball and Cech (1996), who have an excellent survey of these issues. My chapter is concerned with the third of these questions: Would one reach similar conclusions using professionals as opposed to the standard subject pool?[1] However, the papers included in the review might have been motivated by questions of the type 2 highlighted above.

The two categories of subjects—students and professionals—are rather vague concepts. First, who are the typical subjects? The typical subject pool for economic experiments is undergraduate students—in many cases, mostly economics and business undergraduate students (in some cases also graduate students). This is, mostly, a result of the recruitment technology, which used to involve going to large undergraduate classes. Nowadays, with email recruiting, targeting a larger population is feasible; however, it is still the case that the subject pool consists almost exclusively of undergraduate students. At NYU, for instance, we recruit from all fields, but only undergraduate students. Other labs (such as Pittsburgh)

allow nonstudents, but that remains a minority of the subject pool. Recruiting from undergraduate students means that most experiments are composed of a nonrepresentative sample of the population at large in a few dimensions: gender, race, education, (family) wealth, age, and so on. Professionals are loosely defined as people working in an industry where the game under study is thought to be relevant. This definition leaves room for interpretation. For instance, one paper which I hesitated to include is the one by Cadsby and Maynes (1998a), where the game is a public good provision game and the professionals are nurses. Although it is sensible to think that nurses sometimes face situations that have a public good provision structure, it is unclear what makes nurses professionals at public good provision; they are paid for the work they provide, and their wage might internalize all the positive externalities of their services.

The papers included in the review had to meet the following criteria. They had to include both a sample of typical experimental subjects[2] and a sample of subjects who are professionals at the task with which the experiment is involved.[3] They also needed to follow what I will loosely define as standard laboratory procedures. That is, subjects are recruited to participate in an experiment, and thus they know that they are participating in an experiment.[4] Furthermore, subjects are given instructions written using neutral language.[5] Finally, the papers included do not merely compare these two groups; they do so in an environment aimed at testing certain theoretical predictions.[6] Although such narrow focus excludes some interesting studies, it allows us to focus on a very precise question: Do professionals behave differently than the standard experimental subject pool in a typical experimental setting as it pertains to evaluating a certain theory? If they do, it suggests that experimental results are not robust, and to the extent that we care about the behavior of professionals more than that of other groups, this could be a serious problem.[7] If they don't, it suggests that the subject pool per se is not a threat to the external validity of the standard experiment. This is very different, however, from saying that ecological validity is not a concern. These issues will be clarified after concepts relating to the validity of knowledge claims are defined more clearly.

The papers are loosely organized in four thematic sections: other regarding preferences, market experiments, information signals, and a miscellaneous group. The grouping is not important to the analysis; it is only meant to make the review easier to read. Within each section the papers are arranged chronologically. Whenever it was possible, a short description of the typical results for a certain type of experiments is given. Prior to reviewing the papers, I will define some terms and mention issues of interpretation of the results.

Some Methodological Concepts Defined

Since economists only very rarely discuss methodology, I will define a few concepts that will help clarify the relevance of the question asked in this paper.[8] Campbell (1957) first introduced the concepts of internal and external validity, which

were later on expended to four aspects of the validity of an inference (Cook and Campbell, 1979): statistical conclusion validity, internal validity, construct validity, and external validity. To start, it is important to clarify that internal validity has made its way into the experimental economists lexicon with a different meaning. Most (experimental) economists use internal validity to mean the extent to which the environment that generated the data corresponds to the model being tested. This, as will become clear, is much closer to construct validity. For simplicity, in what follows, we will refer to x and y where a claim is made that x causes y.

Statistical conclusion validity has to do with whether the x and y covary, and how strong is that relation. Any factors that increase type I or type II error rates, for instance, have a negative impact on statistical conclusion validity. Misspecification in the estimation of an equation would be an example of a source of this problem.

Internal validity is about the correctness of the interpretation of variation in x *causing* variation in y—that is, whether or not the attribution of causality is warranted. An example of a violation of internal validity would be a case where a within-subject design is used, the order in which the treatment is applied is always the same, and an order effect is confused with a treatment effect.

Construct validity is the ability of the specifics of the experiment to represent the concepts they are intended to capture. For instance, an experiment by Kagel et al. (1995) investigated the impact of wealth on time discounting in rats by having animals with relatively higher and relatively lower weight (weight was controlled by the amount of food the animals were given access to) participate in choices with different rewards and delays. Here the ability to interpret the results as saying that wealthier animals have a high or lower discount factor than poorer ones depends on the validity of representing the construct of wealth by weight and access to food. The use of the term internal validity by economists often corresponds to construct validity as defined here. Experiments that test models are less subject to issues of construct validity insofar as the theorist is often the one choosing the constructs and defining the environment. The experimenter working with a model often simply operationalizes the constructs as defined in the model. That doesn't mean that there are no issues of construct validity, because, for instance, the way the constructs are explained to the subjects matters for construct validity and is an issue independent of whether a model is being tested or not. However, when a theorist writes a model where a firm is modeled as an agent and an experimenter has a subject make decisions to represent that agent, the issue of construct validity seems more acute for the theorist modeling a firm as a single agent than it does for the experimenter using one subject to take the decision of that agent.

Finally, external validity is the ability of the causal relation from x to y to generalize over subjects and environments. External validity does not have to be about generalizing from the subjects or environment in the experiment to subjects and environment outside the laboratory. It can also be about variations in subjects and environments within the experiment (for instance, does the result apply to both men and women in the experiment?).

One concept that economists sometimes blend in with external validity is ecological validity. The two are very different, however. Ecological validity is not of the same nature as the four concepts defined previously, but rather a method where the experiment and its participants are made as similar as possible to what is found in the setting of interest. Hence, an experimenter asking a question relevant to stock trading can design a computer interface exactly like the one at the New York Stock Exchange and recruit professional traders from New York which would make for an experiment with high ecological validity. However, suppose that the causal relationship established between x and y in that experiment did not generalize to non-American traders or to the case when the computer interface from other stock exchanges around the world is used, then that knowledge claim would have little external validity.

Although this chapter speaks to questions of ecological validity, my interest is with respect to the question of external validity.

Some Caveats and Notes

If a model is rejected using undergraduate students while it finds support using professionals because professionals are better at the task at hand, it clearly undermines the external validity of the result, but does it imply that the practice of using undergraduate students is misguided? Models usually do not start with "an agent, who is a professional at doing . . . " The model specifies the environment and the incentives and that is what an experiment, whether it uses undergraduate students or not, recreates. Thus, a failure of the predictions of the model is a failure whether the subjects were undergraduate students or professionals. However, it is true that, to the extent that the model is used to think about the world, and that outside the laboratory it is often professionals who evolve in the environment of interest, then the failure of the model with undergraduate students might be of little interest.

On the other hand, if the predictions of a model are not rejected using undergraduate students but they are using professionals,[9] then, in terms of testing the model, one could argue that the students were a better population than the professionals. All that is indicated by the failure with professionals is that some element that matters in the environment where those professionals evolve is lacking in the model being tested.

One point worth noting, since it influences my take on the papers reviewed, is that I am more interested in the comparative statics and qualitative nature of the results than in tests of point predictions when it comes to testing models. To illustrate what I have in mind, consider the following hypothetical example. Suppose that in an ultimatum game experiment (a proposer offers a split of a pie to a responder who can accept, in which case each party receives the proposed split, or reject, in which case they receive nothing) with students the results are the following. The average offer is 0.45 of the pie, and 95% of offers at or above 0.4 are accepted while 85% of offers below 0.4 are rejected. Furthermore, when the game

is instead a dictator game (a proposer offers a split of a pie to a responder and each party receives the proposed split) then offers decrease to about 0.15 of the pie. To me the interesting results are: the fact that the average offer in the ultimatum game gives a large fraction of the pie to the responder, the fact that nonzero offers are rejected, and the fact that offers are substantially reduced in the dictator game. Suppose that a similar experiment with professional negotiators finds that the average offer is 0.4 of the pie, and 95% of offers at or above 0.35 are accepted while 80% of offers below 0.35 are rejected; also suppose that when the game is instead a dictator game, then offers decrease to about 0.1 of the pie. From those results, I would conclude that the findings are robust, even if all the numbers mentioned above are statistically different for the students and the professionals. Hence, in my review, I sometimes classify the results as the same for the two groups even if there were quantitative differences. To be considered different, the two groups should produce results which lead to a different interpretation of behavior with respect to the model's prediction.

One much earlier paper about issues relating to subject pools is John Kagel's 1987 chapter in *Laboratory Experiments in Economics: Six Points of View* entitled "Economics According to the Rats (and Pigeons too): What Have We Learned and What Can We Hope to Learn?" The introduction to that chapter presents a case for why experiments using animals may be informative for economics. I highly recommend reading that chapter to people interested in the issues discussed in this review (even if they are not interested in animal experiments). For instance, while reading the Kagel chapter, one can easily replace "animal" with "undergraduate student" and "human" with "professional," and many of the arguments will still resonate.

Finally, let me point out that some of the papers covered here did not set out to compare students and professionals, while others are simply interested in different aspects of the results than I am, and thus my description of the results does not always parallel the focus of the papers in question.

Review

Other Regarding Preferences

Bargaining Behavior: A Comparison Between Mature Industrial Personnel and College Students

Fouraker et al. (1962) studied male undergraduates at The Pennsylvania State University (42 subjects) and compare their behavior to that of General Electric employees working in the Industrial Sales Operation division (32 subjects) in a bargaining game that they refer to as a price leadership bilateral monopoly.[10]

The seller (the price leader) picks a price and the buyer then selects a quantity to be traded. A given price–quantity pair implied certain profits to the buyer and seller, such that for a given quantity, a higher price generates more profits for the seller and less for the buyer, and for a given price, a higher quantity first increases profits and then decreases profits for both buyers and sellers. There were no communications besides price and quantity. Prices could vary between 1 and 16, and quantities between 0 and 18, such that the (inefficient) equilibrium involves a price of 9 and a quantity of 10. However, there exists a price–quantity pair (4, 15) which is efficient and involves equal profits for both players (higher for both than in equilibrium). There are two treatments: (a) complete information, where both sides know the profits of each other, and (b) incomplete information, where subjects are only informed of their own profits.

Although the particulars of this game make it different from most bargaining experiments, it seems fair to say that the central tension is the same as in most bilateral bargaining experiment where one side moves first and the equilibrium predicts an advantage to that player, such as in the ultimatum game.[11] In those games, it is frequent to observe divisions of the pie that are much closer to being equal than what is predicted by the subgame perfect equilibrium (Roth, 1995). Although these authors precede the current wave of research on other regarding preferences, their intuition was much in line with it, namely that with incomplete information the equilibrium would emerge (since subjects have no way to know it is not equal nor efficient) while with complete information, results would be closer to equal-split outcome.

Incentives were such that at the end of the experiment, professionals would make almost $49 if they played the equilibrium, and a bit more than $54 if they played the equal-split outcome. Professionals played 3 practice rounds and played 11 rounds for money. The authors separately analyze the first 10 of these rounds as regular rounds, with the 11th regarded as different because it was announced to be the final round. This is relevant in this case as subjects were in fixed pairs throughout the experiment and, thus, aspects of repeated play could have mattered. Students were paid such that if they played the equilibrium they would have made about $5.06 on average, while average profits under the equal split outcome would have been $5.68. They played 3 practice rounds and then played 20 rounds with the same payoffs, with the 20th being announced as the final round, plus a 21st round with triple the payoffs used in the previous 20 rounds. For both professionals and students, they used a between-subjects design. The authors offer no discussion of how the incentives compared to each group's typical earnings.

The results for both groups of subjects (confining attention to the rounds played for money) are that prices start close to the equilibrium, but move downwards over rounds in the complete information treatment, toward the equal-split price. After about half of the regular rounds (which differs across groups), prices have settled at the equilibrium in the incomplete information case and at the equal-split price in the complete information case. Besides the speed of adjustment, the

only main difference is that in the final round, professionals move slightly toward the equilibrium price in the complete information case (about a quarter of the way on average), something which doesn't happen with students.

Overall, results from professionals are in line with results from students, namely they display a tendency toward outcomes that equalize payoff when they have the information allowing them to do so, and toward equilibrium otherwise. However, in the final round, professionals did show a tendency for strategic behavior that the students did not display. Unfortunately, the changes in protocol between the two studies make it difficult to draw strong conclusions one way or the other.

Choosing Between a Socially Efficient and Free-Riding Equilibrium: Nurses Versus Economics and Business Students

Cadsby and Maynes (1998a) studied undergraduate and graduate economics and business students at York University and the University of Guelph (6 sessions of 10 subjects labeled experiments E1–E6) and compared their behavior to that of registered nurses at four hospitals and one college (6 sessions of 10 subjects labeled experiments N1–N6) in a threshold public goods game.

At the beginning of each period, each participant receives 10 tokens. They can contribute any portion to a group fund. If 25 or more tokens are put in the group fund, each player receives 5 additional tokens. Payoffs are 10 – *contribution* if the threshold is not met, and 15 – *contribution* if the threshold is met. This is repeated 25 times.

There are two pure strategy Nash equilibria: Either each player contributes 0, or the group contributes just enough to meet the threshold (which is Pareto optimal). There are an infinite number of mixed strategy equilibria.

It is not clear what is the standard result in this environment. Croson and Marks (2000) report, in their meta analysis, three other studies with a similar design (threshold public goods game with no refund, no rebate, no communication, no heterogeneity in endowments or valuation, and a step return of 2 – the ratio of the aggregate group payoff of the public good to the total contribution threshold); the differences are the number of players, the endowments, and the subject pools (Canada or the United States). In the paper by Cadsby and Maynes (1998b), contributions go toward 0, whereas in the papers by Marks and Croson (1998) and Croson and Marks (2001) contributions go to the efficient level. According to the meta study, this divergence in results would result from the different endowment since the impact of the number of players is supposed to be in the opposite direction (the country where the study was conducted is not considered as a possible factor).

The incentives were the same for both types of subjects: 12 cents per token. Thus, if they were paid the average of the two equilibria times 25, they would make $31.56, which was about the opportunity cost for nurses for 1 hour and a half.

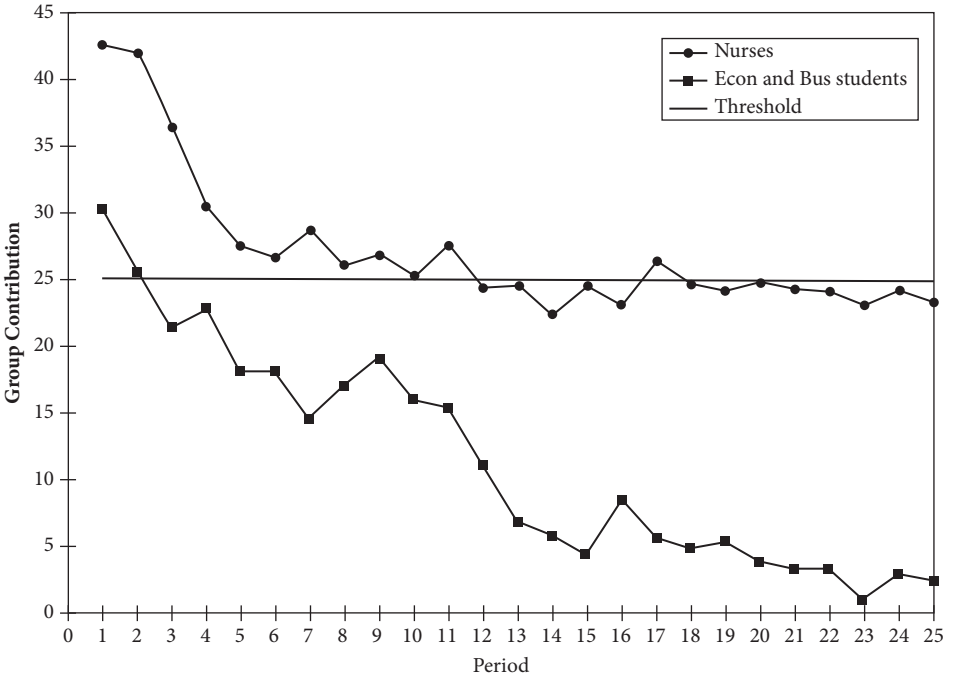


Figure 17.1. Average group contribution by period: nurses versus economics and business students.

The results are that nurses started well above the threshold and finished close to the threshold. On the other hand, students started close to the threshold and finished far below the threshold. The average contributions are plotted in Figure 17.1. In the last 5 periods, the authors report that neither student nor nurses are closer to a NE equilibrium; that is, the distance between their contributions and the closest NE is not statistically different.

Thus, it seems that nurses contributed more. However, the frequency with which they provide the public good, albeit being higher than for students, is not statistically significant. This raises the question, Would these differences persist if the experiment had lasted longer? In the face of the threshold not being met, one can wonder how much longer the nurses would have continued contributing.

Another question left open is, How would a nurse in a group of students have behaved, and how would a student in a group of nurses have behaved? Is the difference driven by perceptions (or expectations) of others or by intrinsic differences? The downward trend seems to indicate that it is about perceptions (which they continually adjust).

Nonetheless, this does suggest that subject pools may matter for inferences about coordination situations. On the other hand, neither group is closer to the theory's prediction than the other.

The Hidden Costs and Returns of Incentives—Trust and Trustworthiness Among CEOs

Fehr and List (2004) studied students from the University of Costa Rica (126 subjects) and CEOs from the coffee mill sector (76 subjects) in a trust game.

Two versions of the game are studied. In both there is a principal and an agent and they are paired anonymously. Both receive 10 experimental currency units (ECUs). In the trust treatment, the principal transfers $x \in \{0, 1, 2, \ldots, 10\}$ and announces a desired back transfer $\hat{y}$. Then, the agent receives $3x$ and returns $y \in \{0, 1, 2, \ldots, 3x\}$. In the trust with punishment (TWP) treatment, in addition to x and $\hat{y}$, the principal chooses if he wants to impose a fine of 4 if $y < \hat{y}$. The agent is told x, $\hat{y}$, and f before choosing y.

In both treatments, $x = y = 0$ is an equilibrium. In the TWP treatment there are other equilibria such that $f = 4$ and $y = 3$ or 4.

The incentives were 1 ECU converted to $0.20 for students. Average earnings were $5.65 for 45–60 minutes. The typical earnings are $2 per hour. Professionals faced a conversion rate of 1 ECU equal $2.00. This resulted in average earnings of $65 for 45–60 minutes.

The typical results in trust games are that senders send money and responders send back money. However, it is often the case that, on average, responders send slightly less than senders have sent on average (Berg et al., 1995).

The results that apply to both groups are the following. Principals choose $x > 0$ on average (in both treatments). Principals choose $3x > \hat{y} > 0$ on average (in both treatments). Agents send money back $3x > \hat{y} > y > 0$ on average (in both treatments). The effect of the treatment manipulation is negligible. More principals opt to use $f > 0$ than not. Payback is higher if $f = 0$. The differences between students and professionals are that CEO principals transfer more money than students. Also, CEO agents send back more money than students. Finally, CEO principals use the $f > 0$ less than student principals.

In conclusion, the main result seems to be that the insights that one gains from the trust game experiments (and TWP) using students are supported by experiments with CEOs.

Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyama Bay

Carpenter and Seki (2005) compare university students from Japan (26 students –2 sessions) to Japanese shrimp fishermen (27 subjects –2 sessions) in voluntary contribution mechanisms experiments. In their daily fishing activities, some of the fishermen are organized in a group that shares both income and operating expenses (referred to as poolers) and a group that doesn't (nonpoolers).

The experiment has two parts for a total of 10 rounds. Throughout they are divided in groups of 4 (since the group size were not multiples of 4, some subjects were "active" in more than 1 group). In the first part of five rounds, subjects do

a standard voluntary contribution mechanism game. That is, each subject is endowed with 500 yen. They can keep the money for themselves or contribute to a public good. Everything contributed to the public good is doubled and divided equally by the members of the group. This was followed by five rounds of the social disapproval protocol. The social disapproval protocol is a modification of the standard VCM in which after subjects have decided on their contributions, they are told each member's contributions and asked if they want to send an unhappy face to the rest if the group at a cost of 10 yen to themselves. That signal is displayed to the group at the beginning of the following round.

The subgame perfect equilibrium of both games (with standard preferences) implies no contributions. In the case of the social disapproval protocol, it implies no signals.

The typical result in experiments on the VCM games can be summarized as follows: Positive contributions to start, but contributions decrease to relatively low levels over repetitions of the game, and settle slightly above 0 (Ledyard, 1995).

The experiments typically lasted less than one hour. Average earnings were $73.19.

Combining all 10 periods, poolers contribute slightly more than nonpoolers, and together they cooperate significantly more than students. However, in the standard VCM, poolers and nonpoolers are not statistically different, while post disapproval rates are significantly different. Social disapproval rates are lowest for poolers and highest for students, but of course this is partly driven by the different contribution levels. Going from the standard VCM to the social disapproval protocol, contributions increase for poolers while they decrease for both nonpoolers and students.

Thus, contribution levels vary across subject pools, including across professional subject pools. The direction of change across treatments is the same for students and nonpoolers but is different for poolers.

Market Experiments

The two experiments under this heading are different in their focus. The first one is a variation on the standard oral double auction (ODA). The other could have been categorized differently, it studies a principal–agent problem within a sealed-bid market structure (sellers make offers). However, they have been grouped together due to the extreme power of market environments to bring about the "correct" outcomes, in the case of the ODA even in the face of extremely limited agents (Gode and Sunder, 1993). One could say that under most conditions, ODA quickly results in equilibrium outcomes (Holt, 1995). However, "posted-offer markets tend to competitive predictions more slowly than comparable DA and they tend to converge to the competitive price prediction from above (if at all), implying a transfer of surplus from consumers to producers in the adjustment process" (Davis and Holt, 1993, p. 217).

Experience and Decision Making: A Comparison of Students and Businessmen in a Simulated Progressive Auction

Burns (1985) compared the behavior of second-year microeconomics undergraduates (who were part of an economic course) and experienced wool buyers (senior buyers with an average 35 years of experience) in a progressive oral auction with homogeneous commodities.

There are two groups of nine student subjects and one group of nine professional subjects (four are officers of the Australian Wool Corporation and five represent the leading wool brokerage houses.) The progressive oral auction with homogeneous commodities is close to the market structure faced by wool buyers.

In each market every buyer has a demand for 2 units, the second having lower returns than the first. There is a supply of 12 units (unknown to the bidders). Every bid must be an improvement bid. When no bid has been placed for 5 seconds, the item is declared sold. There are penalties for untraded units, which is a way to stimulate trade and is argued to represent the fact that the requirement to meet the full demand quota is very serious in the wool market.

Fifteen auctions are conducted in total (5 per week). Conditions are constant within weeks but demand changed across weeks. Traders are informed of the direction, but not the magnitude, of the change.

The standard result for oral double auction experiments is that price and quantity traded converge rapidly to equilibrium predictions (Holt, 1995). Hence comparative static predictions of the impact of demand and supply change on the direction of change of the quantity and price variables is as predicted.

Table 17.1 is taken from the paper and indicates the values for each bidder and the predicted market prices. The key prediction of interest is that under perfect competition, the market clearing prices would be 96.5, 90.5, and 98.5 in weeks 1, 2, and 3, respectively.

The incentives for the students were very unusual and rather vague. The experiment was part of a course exercise for which an essay worth 10% of the students final assessment must be written. The students do not know the subject of the essay, but they are advised that "only by striving to maximize their profits would they gain the understanding necessary to successfully complete the assignment." The incentives for the professionals are also very unusual. It was announced to the wool buyers that "the best trader would be revealed at the end of the session."

The results can be summarized as follows. Wool buyers bid up to their marginal values on the first lot, then marginal value plus penalty on the second. Students behave similar to wool buyer on day 1 of week 1, and then the curve flattens out in subsequent days (more contracts at or close to the market equilibrium prediction). In each successive week the period of adjustment to profit maximizing level of prices grows shorter and they slowly incorporate the penalty in their bids over time. As a result, students make much more money than wool buyers. From one week to another, both students' and wool buyers' change in behavior is such that price adjusts in the right direction.

Table 17.1. Buyers' Valuations[a]

	Week 1			Week 2			Week 3		
	Lot 1	Lot 2	Av.	Lot 1	Lot 2	Av.	Lot 1	Lot 2	Av.
Buyer	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	115	96	105.5	96	93	9405	114	97	105.5
2	113	95	104	95	92	93.5	111	96	103.5
3	112	95	103.5	94	91	92.5	109	95	102
4	109	94	102.5	109	90	99.5	106	95	100.5
5	107	93	100	107	89	98	104	101	102.5
6	104	93	98.5	106	89	97.5	103	100	101.5
7	102	99	100.5	103	88	95.5	102	99	100.5
8	101	98	99.5	101	87	94	117	98	107.5
9	101	97	99	98	87	92.5	115	97	106
Expected average market price (= equilibrium value) if marginal values are used.			96.5			90.5			98.5
Expected average market price (= equilibrium value of column 3) if average values are used.			99.75			93.75			101.75
Expected average market price (= average of 2nd to 13th valuations) if full-value bidding is used.			103.25			97.25			105.25

Throughout this table, Australian dollars are used as the monetary unit.
Source: Holt (1995).

Discussions and interviews with the wool buyers suggest that their behavior was driven by the behavior they have learned in the market they know. More specifically, one aspect of their professional experience that seems to conflict with the specifics of the experiment is the fact that the goods they deal with in their everyday experience are homogeneous. That, in particular, means that these traders are not accustomed to noticing within-day price variations as these can represent different quality wool. Prices at which wool trades are noted by a junior, and traders receive changed instructions from the head office across days; those changes reflect price trends. As a result, despite the fact that each auction featured a sharp decline in prices in the course of the session, seven of the nine professional buyers reported not noticing that pattern. Another practice that stems from their experience is to bid on units for which they are not interested in to "keep others honest."

A key question left open is whether the professionals would have generated higher earnings if they had been paid; that is, would they have relied less on the reflexes they have developed in the usual environment and paid more attention to the specific details of the new environment they were confronted with?

A Note on the Use of Businessmen as Subjects in Sealed Offer Markets

Dejong et al. (1988) study the behavior of students from the College of Business at the University of Iowa (one session of seven subjects) and of members of the Professional Accounting Council of the Department of Accounting at the University of Iowa (one session with five partners in public accounting or auditing and two corporate financial officers) in an elementary principal–agent problem.

Buyers and sellers are provided with an initial endowment of $1 for the buyers and $6 for the sellers. Each seller submits a sealed-bid offer to buyers that specifies a quality of service and a price. Buyers decide which offer to accept. Then a random draw is performed to determine if the buyer incurs a loss or not given the quality of service he has bought. The parameters were such that the payoffs are as presented in Table 17.2.

The due care standard is set as 3. That is, the seller is responsible for a loss if he provided a level of quality lower than 3.

In a perfectly competitive market the predicted outcome is service of quality 3 provided at a price of $0.26.

Table 17.2. Parameter Values for Markets

Level of Service Quality x	Probability of Loss $P(x)$	Expected Loss[a] $P(x)l$	Change in Prob. of Loss $\Delta P(x)$	Change in Expected Loss $-\Delta P(x)l$	Cost of Service $C(x)$	Change in Cost $\Delta C(x)$	Seller Expected Cost[b]	Expected Social Cost[c]
No purchase	0.90	72[c]			0[c]		0[c]	72[c]
1[d]	0.60	48[c]	−0.30	24[c]	5[c]	5[c]	53[c]	53[c]
3[e]	0.20	16[c]	−0.40	32[c]	26[c]	21[c]	26c[e]	42c[e]
5	0.05	4[c]	−0.15	12[c]	53[c]	27[c]	53[c]	57[c]

[a] Loss(ℓ) = 80c throughout.
[b] Seller's expected cost when the seller bears liability for the loss is predicted to be $P(x)\ell + C(x)$.
Seller's expected cost when the buyer bears liability for the loss is predicted to be $C(x)$.
Under the assumption of risk neutrality, the fee, r, in a competitive market is predicted to equal the expected cost.
[c] Expected social cost if $P(x)\ell + C(x)$.
[d] Seller bears the loss under the negligence liability rule.
[e] Predicted equilibrium price and quality of the service. This quality of service is also the due care standard and level of service that minimizes expected social costs.

Students were paid the sum of their earnings, which ranged between $10 and $25. Professionals faced a different payment structure. First, each professional subject was given an envelope with the average earnings of the student who had the same role as they have. They played for points, at the end, if they had made more points per period than the student they were paired with, they received a University of Iowa Pewter souvenir, otherwise they received nothing.

The results indicate that the prices for each quality of service are not statistically different for students and professionals. As for the quality of service, students met or exceeded the due care standard more often than professionals; however, that difference is not statistically significant. The same conclusions can be reached looking at average expected profits; that is, they are not statistically different for professionals and students (only one of the 18 comparison they make is statistically different). Similarly, out of the 18 comparisons the authors conducted relating to allocative efficiency, only two are statistically different.

Hence, behavior of students and businessman is very similar in this experiment; but just as with the previous study, it is impossible to tell what, if any, role the unusual (and different for the two groups) incentives played.

Information Signals

This section includes two signaling experiments and one experiment about information cascades. All these models rely on Bayesian updating, something which, in general, humans have shown not to be extremely good at (Camerer, 1995). However, models of signaling games have found support when subjects are given ample experience (Cooper et al., 1997), and studies on information cascades also revealed that many of the model's predictions are borne out in the data (Anderson and Holt, 1997).

Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers

Cooper et al. (1999) compare college students from China Textile University (10 sessions) to managers (12 sessions). Managers include older managers (high-ranking managers, mid-level managers, and senior foremen) from textile factories in Shanghai (10 sessions), younger managers who are graduate students at China Textile University and had spent at least 5 years in factories before returning to school for an MBA-type degree, and China Textile University alumni working in the area with at least 2 years' working experience (two sessions). They compared the behavior of these groups in the ratchet effect game.

The game played in the lab is a simplified version that generates a standard signaling game which is thought to represent a typical problem in centrally planned economies, namely the fact that production targets assigned by the central planner to specific firms increase with productivity, thus giving an incentive to firms to

Table 17.3. Firm Payoff and Planner Payoff

Firm Payoff

	Low-Productivity Firm Production Target		High-Productivity Firm Production Target		
Output	EASY	TOUGH	EASY	TOUGH	Output
1	710	542	1,108	815	1
2	730	561	1,145	852	2
3	697	528	1,230	937	3
4	527	357	1,308	1,015	4
5	273	103	1,328	1,035	5
6	220	48	1,298	1,005	6
7	190	15	1,250	966	7

Planner Payoff

Production Target	Facing Low-Productivity Firm	Facing High-Productivity Firm
EASY	645	764
TOUGH	528	820

misrepresent their true productivity. Prior to the start, firms learn their type (50% probability of each). Firms choose output; and at any output, payoffs are higher if the central planner chooses "easy" rather than "tough." Central planners see the output level chosen and select "easy" or "tough". The central planner's payoffs are a function of the target and the firm type. The payoffs are indicated in Table 17.3.

About half the sessions were conducted in generic terms and half were in context (referring to easy and tough contracts, high-productivity firms, etc.). Each session had 12 to 16 subjects. The game was repeated 36 times with role reversed after every 6 games.

The game has three pure strategy sequential equilibria: pooling at output levels 1, 2, or 3.

The incentives were the following. The students had two payment schedules. In the standard pay cases, pooling at 2 corresponds to an expected payoff of 30 yuan ($3.75), which is about equivalent to the earnings in a typical U.S. experiment. In the high-pay cases, payments were multiplied by 5, which represents 150 yuan in the pooling equilibrium at 2. As a point of comparison, the monthly wage for an associate professor was 1200 yuan. Managers had the same incentives as the students in the high-pay condition. Note that the vast majority of managers in the experiment earned less than 2000 yuan per month.

The results are the following. First, we look at all treatments as a whole. Firms' choices start clustered around their type's full information output. Central planners give easier contracts to outputs 1–3 than to higher ones. Experience increases

the frequency of 1–3 outputs by high firms (strategic play) with play converging to 2. However, a sizable frequency of nonstrategic play by high firms remains even in the last 12 games. Second, we focus on students only and the impact of incentive variation. Increased pay promoted more strategic play by firms initially. However, by the end there is no difference. Increased pay has no impact on the central planner's choice. Third, we focus on the impact of the variation in context. There are no effects on students acting as firms. Mistakes by central planners are reduced for students but only in standard pay. There is an increased level of strategic play for managers in their role as firms in later cycles of play. There is a strong effect on managers as central planners, promoting higher target rate differentials than in the generic sessions.

To conclude, similar behavior is observed between students and managers. However, context helps managers.

Professionals and Students in a Lobbying Experiment Professional Rules of Conduct and Subject Surrogacy

Potters and van Winden (2000) compared the behavior of undergraduate students (mostly economics majors) from the University of Amsterdam (12 sessions) and public affairs and public relations officers from both the private and public sector (3 sessions) in a lobbying game.

The game is a signaling game with a sender (S) and a receiver (R). R decides between B1 and B2, and the payoffs of S and R depend on the realization of a stochastic variable (k) which can take 2 values (w with $p = \frac{2}{3}$ and b with $p = \frac{1}{3}$). S sees k and decides to send a message or not. Sending a message costs c to S. After R is informed of S's decision, he picks B1 or B2. Two treatments: low message cost (L) and high message cost (H). The specific payoffs are shown in Table 17.4.

Each session had 10–12 subjects and lasted for 10 rounds of play.

Table 17.4. Firm Payoff and Planner Payoffs (in Dutch Guilders) to Sender and Receiver in the Two Treatments Depending on the Color of the Disk (White or Black) and the Choice of the Receiver (B1 or B2)

		R's Choice	
Payoff to S,R		B1	B2
Low message cost ($c = 0.5$)			
State	White ($P = 2/3$)	2, 3	4, 1
k	Black ($P = 1/3$)	1, 0	7, 1
High message cost ($c = 1.5$)			
State	White ($P = 2/3$)	1.5, 3	3.5, 1
k	Black ($P = 1/3$)	1.5, 0	5.5, 1

The game has a unique sequential equilibrium that satisfies "all" the refinements. In treatment L: If $k = w$, S sends a message with $p = \frac{1}{4}$; otherwise he sends a message for sure. If R gets no message, he picks B1; otherwise he picks B2 with $p = \frac{1}{4}$. In treatment H: If $k = w$, S sends a message with $p = \frac{1}{4}$; otherwise he sends a message for sure. If R gets no message, he picks B1; otherwise he picks B2 with $p = \frac{3}{4}$.

The incentives for the students were such that the expected equilibrium earnings were 20 guilders per hour (typical earnings are 15 guilders per hour). The professionals received 4 times as much, hence in equilibrium that represents 80 guilders per hour (approximately their estimated wage per hour).

The results that apply to both subject pools are the following. S engages in costly signaling. R responds in the expected direction. R chooses B2 more after a message in the H treatment than in the L treatment. More messages are sent when $k = w$ in the L treatment than in the H treatment (not predicted by the theory). Two differences emerge between the two groups. First, the frequencies with which professionals send messages following w versus b are closer to the equilibrium predictions than for the students. Second, professionals in the role as S earn slightly more than students (it is not clear from the paper if that is significant).

To conclude, the results indicate only minor substantive differences. The core result seems to be that students, just like professionals, understand the central strategic tension in this game.

Information Cascades: Evidence from a Field Experiment with Financial Market Professionals

Alevy et al. (2009) compare students at the University of Maryland (10 sessions with 54 subjects) and market professionals from the floor of the Chicago Board of Trade (10 sessions with 55 subjects) in a game about information cascade formation.

The game is the following. The state of nature (A or B) is selected with the roll of a die, but not revealed. Subjects draw a signal from an urn. Subjects make their choice sequentially with their choice revealed to others. In the symmetric treatment, urn A contains two type a balls and one type b ball; urn B contains 2 type b balls and 1 type a ball. In the asymmetric treatment, four type a balls are added to both urns. The other treatment variation is that some treatments are framed as gains and others as losses (gain for correct guess or loss for incorrect guess). This is repeated for 15 rounds in groups of 5 or 6.

The theoretical predictions for the state of the world being A are presented in Table 17.5.

Students received \$1 per correct guess or –\$1 per incorrect guess. Professionals received \$4 per correct guess or –\$4 per incorrect guess. which made the median payout greater than \$30 for 30 minutes.

For both subject pools, a majority of choices are consistent with Bayesian updating, and a non-negligible fraction of cascades are realized. Earnings and the rate

Table 17.5. Posterior Probabilities for All Possible Sequence of Draws for Both Symmetric (Upper) and Asymmetric (lower, italic) Treatments Based on Choice Histories (*a*, *b*)

Posterior Probabilities: Symmetric (upper) and Asymmetric (lower) Urns								
a	*b*	0	1	2	3	4	5	6
0		0.500	0.330	0.200	0.110	0.060	0.030	0.020
		0.500	*0.330*	*0.200*	*0.110*	*0.060*	*0.030*	*0.020*
1		0.670	0.500	0.330	0.200	0.110	0.060	
		0.545	*0.375*	*0.231*	*0.130*	*0.070*	*0.036*	
2		0.800	0.670	0.500	0.330	0.200		
		0.590	***0.419***	*0.265*	*0.153*	*0.083*		
3		0.890	0.800	0.670	0.500			
		0.633	***0.464***	***0.302***	*0.178*			
4		0.940	0.890	0.800				
		0.675	*0.509*	***0.341***				
5		0.970	0.940					
		0.703	*0.554*					
6		0.980						
		0.749						

Note: The prior probability of an urn is 0.5 in (0, 0). Bold entries for the asymmetric urn are those in which counting and the posterior probability make different predictions about the state.

of cascades are not different across subject pools. The differences are the following: Professionals are slightly less Bayesian than students; professionals enter in fewer reverse cascades; and, finally, although students' Bayesian behavior is affected by the gain and loss domain, the professionals' Bayesian behavior is not.

Thus, although there are differences in the details, the main insights from the student population extend to the professionals.

Other Topics

A Comparison of Naive and Experience Bidders in Common Value Offer Auctions: A Laboratory Analysis

Dyer et al. (1989) compare undergraduate students from upper-level economics majors at the University of Houston (three sessions labeled experiments 1–3) to executives from local construction companies (one session labeled experiment 4) in a series of sealed-bid first-price common-value auctions. The professionals have, on average, 20 years of experience in the construction industry—all but one have many years' experience in bid preparation.

The game is the following. Subjects have the right to supply a single unit of a commodity, and the commodity is awarded to the lowest bidder using a sealed-bid first-price auction. The cost, C, is unknown at the time of the bid. The winner earns $bid - C$ while the others earn 0. C is drawn from a uniform on [\$50, \$250]. Each bidder receives a private signal c_i drawn from a uniform on $[C - \varepsilon, C + \varepsilon]$. Distributions for C, c_i, ε, and N are common knowledge.

Experiments 1–4 start with four active bidders. In experiment 4, after 24 periods 1 professional and 2 "experienced" students subjects are added both with and without an announcement of the highest cost estimate C_H.

Theory predicts that for costs in the interval $[\$50 + \varepsilon < c_i < \$250 - \varepsilon]$, the symmetric risk-neutral Nash equilibrium (SRNNE) is

$$b(c_i) = c_i + \varepsilon - Y,$$

where $Y = \left[\frac{2\varepsilon}{N+1}\right] \exp\left[-\left(\frac{N}{2\varepsilon}\right)(250 - \varepsilon - c_i)\right]$. Y diminishes rapidly as c_i moves below $\$250 - \varepsilon$. Thus, in equilibrium, bids mark up signals by close to ε.

The incentives are the same for both types of subjects (although the number of periods differ and the details of the auction are varied for part of experiment 4). The starting balance is \$10. Losses and profits from every auction played are added to the starting balance. Thus under the SRNNE this represented about \$100 in experiment 4, which lasted about 3 hours.

The result that really stands out in the experimental literature on common-value auctions is the failure to account for the winner's curse and thus the persistent finding that bids are such that losses would result on average. One result which is somewhat of a hallmark of the winner's curse is the fact that many experiments show that losses increase as the number of bidders is increased (Kagel and Levin, 2002).

In their experience, the authors observe that for executives, negative or near-zero profits dominate, there is little evidence of learning over time, and losses increase as N increases, announcing that C_H raises the offer price (contrary to the theoretical prediction).

Comparing the executives to the students, they find that both fall prey to the winner's curse. There are no differences (at 10% level) for any of the following: percentage of times the low bid is submitted by the low signal holder, average actual profits, and percentage of times the low bid is such that it would result in losses on average. There are small differences in terms of the impact of changes in ε. There are also differences in the details of the reaction to changes in N as compared to other experiments with students. The impact of C_H is consistent with prior experiments with students.

Thus, executives and students' behavior are similar in that both exhibit the winner's curse and are similar on many other relevant dimensions. However, there are some differences in the details. One question this raises is, How are these executives successful in their business? Based on discussions with the executives, the

authors conclude that the executives have learned a set of situation-specific rules of thumb which permit us to avoid the winner's curse in the field but could not be applied in the lab. This is explored in more detail in Dyer and Kagel (1996).

Experientia Docet: Professionals Play Minimax in Laboratory Experiments

Palacios-Huerta and Volij (2006) studied how non-economics and non-mathematics majoring male Spanish university students (160 students) and male professional soccer players (40 kickers and 40 goalkeepers) behave in two zero-sum games.[12]

The first game is the 2 × 2 zero-sum game in Table 17.6.

The numbers used correspond to the probabilities that a penalty kick would be successful. However, no soccer terminology or reference was made in the experiment. Subjects played 15 rounds for practice and 150 rounds for money, at 1 euro per round. Experiments lasted, on average, 70 minutes for professionals and 81 minutes for students. The second game is taken from O'Neil (1987). It corresponds to the zero-sum game in Table 17.7.

The payoffs are in Euros. This was played 15 rounds for practice and 200 rounds for money. On average, sessions lasted 61 minutes for professionals and 68 minutes for students.

The unique mixed strategy equilibrium of the penalty kick game is for the kicker to play Left with probability 0.3636 while the goalkeeper plays Left with probability 0.4545. The unique mixed strategy equilibrium of O'Neil's game is for both players to play Green with probability 0.4 and to play each of the other choices with probability 0.2.

Table 17.6. Penalty Kick Game

		Goalkeeper	
		Left	Right
Kicker	Left	0.6, 0.4	0.95, 0.05
	Right	0.9, 0.1	0.7, 0.3

Table 17.7. O'Neil Game

		2			
		Red	Brown	Purple	Green
1	Red	0, 1	1, 0	1, 0	0, 1
	Brown	1, 0	0, 1	1, 0	0, 1
	Purple	1, 0	1, 0	0, 1	0, 1
	Green	0, 1	0, 1	0, 1	1, 0

The results are clearly supportive of the minmax model for the soccer players. Namely, choice frequencies in both games are either (a) not statistically different from that predicted by the model or (b) so close that the difference is not important. Furthermore, the data support the implication of the theory that choices are independent draws. On the other hand, the results for students are much further away from the theory. In the penalty kick game, the aggregate frequencies are not too far off; however, there aren't enough differences in the choice frequencies of kickers and goalkeepers. Furthermore, unlike soccer players, students do not generate "random" sequences. Similarly for O'Neal's game, students aggregate behavior is similar to the model's predictions, but not after one looks at individual behavior.

Non-economics and non-mathematics majoring male Spanish university students do not behave the same as male professional soccer players in the zero-sum games studied.

Option Pricing by Students and Professional Traders: A Behavioral Investigation

Abbink and Rockenbach (2006) compare students (mostly from the economics and law departments) from the University of Bonn (108 students in six treatments) to 24 employees (two treatments) from"an influential German Bank in Frankfurt/Main" who were "decision makers" in their "departments of foreign exchange, security, futures, bonds, and money trade" (p. 503) in an option pricing (individual decision) experiment. They also add an additional two treatments of students from a different institution: the University of Erfurt, which has mainly social science students in nontechnical fields. In particular, option pricing is not part of the curriculum in any courses offered at the University of Erfurt, while it is part of the Finance module, a popular one with economics students, at the University of Bonn.

Subjects have 600 units of experimental currency to invest in one of three potential forms, namely, two risky investments, X and Y, and a riskless investment which is to keep the currency (at an interest rate of 10%). The returns to the risky investments depend on the state of the world tomorrow: *Red* (with probability q) or *Blue* (with probability $1 - q$). The subjects are first offered to buy or sell up to 9 units of X at price P. Following that, they can buy, sell, or not trade in Y at a price of 100 per unit with the constraints that they can only trade up to a quantity such that they will not be bankrupt tomorrow. The cash flows per unit traded are represented in Table 17.8.

A rational investor in this environment follows an option pricing strategy with a separating price of 20. That is, the investor buys X when P is less than 20 and sells it when it is above 20. This result is independent of q, the probability of states *Red* and *Blue*.

The experiments used a between-subjects design varying q across values 10, 20, 33, 50, 70, and 90 for students and used values 20 and 70 for professionals. Students subjects played 50 rounds, and in each round a different price P between 2 and 62

Table 17.8. The Investment Possibilities[a]

Cash Flows of the Investment Forms							
	When Selling				When Buying		
	Today	*Red*	*Blue*		Today	*Red*	*Blue*
X	+*P*	−62	−2	*X*	−*P*	+62	+2
Y	+100	−150	−90	*Y*	−100	+150	+90

[a]The interest rate for cash is 10%.

was randomly selected. The professionals played 30 rounds with prices drawn from the odd numbers between 2 and 62.

The experimental currency was exchanged for money at an unspecified rate. The student sessions lasted between 1 and 2 hours whereas the professional sessions lasted 1 hour. Abbink and Rockenbach (2006, pp. 503–504) say the following about the exchange rate for professionals: "The exchange rate of the thalers [the experimental currency] earned in the experiment was chosen such that it was comparable to the one used in the corresponding students treatment."

Four results stand out. First, students react more to q than professionals do; that is, the separating price that best fits the subjects behavior varies more with q. Remember that in theory, q should not affect the separating price. Second, the estimated average separating price is closer to the theoretical one for students than for professionals. Third, over time, students move closer to the theoretically predicted price, whereas the professionals move in the opposite direction. A subset of the results (for the cases where data are available for students and professionals) are reproduced in Table 17.9. Fourth, the professional traders exploit arbitrage opportunities less frequently than students, and as a consequence their behavior implies lower expected value exploitation.

In the additional treatments conducted with students from the University of Erfurt (using the same protocol and treatment parameters as for professionals), it was found that their behavior is closer to that of professionals.

As a whole, it seems that professionals behavior is further from the predictions of the theory than that of students, although students fail more severely in terms of their reaction to the treatment variable q.

Table 17.9. The Separating Prices Over Time

	Students		Professionals	
	Treatment, q (%)			
Experience Phase	20	70	20	70
Rounds 1–15	25.17	38.70	27.48	36.49
Rounds 16–30	18.64	35.89	32.08	39.35

Are Experienced Managers Experts at Overcoming Coordination Failure?

Cooper (2006) compares 20 Case Western Reserve University undergraduate students to 19 students of the executive MBA program of that university's Weatherhead School of Management as managers in what he calls a Turnaround game. That is, the managers set bonuses in a weak-link game (also known as a minimum game). The difference between the two pools is covered in detail in the paper, but some of the key facts about the professionals are that they have at least 10 years of relevant work experience, with at least 5 years involving substantial managerial responsibilities. They come mostly from manufacturing, but there are also many in the healthcare and service sectors and their average income is $122,000.

Subjects are in groups of five, one manager and four employees. The employees are all undergraduate students, whereas the managers vary with treatment (professionals or students). For 10 rounds, employees play the minimum game shown in Table 17.10. Managers and employees both observe the minimum of the group choice. In the following 20 rounds, managers select a bonus (B can be any integer between 6 and 15) and can send a message to the employees before the employees select their effort (E). The bonus is costly to the manager but transforms the game of the employees, and the payoff functions are

$$\text{Manager payoff} = 100 + \left(\left(60 - 4B \times \min_{i\in\{1,2,3,4\}} (E_i)\right)\right),$$

$$\text{Employee i payoff} = 200 - 5E_i + \left(B \times \min_{i\in\{1,2,3,4\}} (E_i)\right).$$

Table 17.11 shows the employees' stage game for a bonus of 14.

The incentives are the same for both groups of subjects. Payoffs are in ECUs and converted to dollars at a 500:1 ratio. Hence average payoffs were $23.89, including the $10 show-up fee. Furthermore, participants were told that a list of earnings ranked from top to bottom would be sent to them. A few aspects of the design that are atypical are the following. First, the managers were separated from the employees, but in particular the professional managers were participating from home over the internet. Second, the instructions changed slightly over time to deal with the fact that professionals were having difficulty operating the software. Third, the instructions were neutral but used words that might be more evocative than what is

Table 17.10. Employee's Payoff When the Bonus Is 6

		Minimum Effort by Other Employees				
		0	10	20	30	40
Effort By Employee *i*	0	200	200	200	200	200
	10	150	210	210	210	210
	20	100	160	220	220	220
	30	50	110	170	230	230
	40	0	60	120	180	240

Table 17.11. Employee's Payoff When the Bonus Is 14

		Minimum Effort by Other Employees				
		0	10	20	30	40
Effort By Employee *i*	0	200	200	200	200	200
	10	150	290	290	290	290
	20	100	240	380	380	380
	30	50	190	330	470	470
	40	0	140	280	420	560

typical (which the author refers to as naturalistic) For instance, employees were told that they had to choose hours per week of work.

The results can be summarized as follows. By round 10, effort is very low (a typical result in minimum games; see for instance, Van Huyck et al. (1990)), averaging 3.08, and is exactly 0 in 31 of the 39 groups. Bonuses and communication allow both groups to increase profits and efforts by coordinating on higher effort levels. The increase in effort is about fivefold between round 10 and round 30, and the increase in profit is about threefold. Over the last 10 rounds, average minimum efforts are essentially indistinguishable between undergraduates and professionals, and so are bonuses and profits. Even more so if one focuses on groups which started at a minimum effort level of 0 in round 10.

Some differences emerge, however. First, professionals manage to increase effort and profits much quicker than undergraduates do. From the graphs in the paper, it looks like it takes about five rounds for undergraduates to increase profits close to the levels of professionals. Second, professionals communicate different messages than undergraduates. In particular, professionals are much more likely to request a specific effort level, they are less likely to lay out a long-term plan, and they are more likely to offer encouragement. They are also much more likely to refer to trust.

Hence, one would have reached similar conclusions using professionals and undergraduates in this experiment in the sense that both learned to use bonuses and communication to improve profits by increasing efforts.

Discussion

Thus, there are 13 papers that allow us to compare students and professionals in a standard laboratory environment (papers are referred to by the initials of its authors). The results are summarized in Table 17.12. In nine of those 13, professionals are not closer or further from the theory in a way that would lead us to draw different conclusions. In the remaining four, only one finds behavior on the part of professionals which is substantially closer to the prediction of the theory,

Table 17.12. Summary of the Distance to the Theoretical Prediction

	Other Register	Market	Signals	Other	Total
Pros closer				PHV	1
Similar	SH FL (CM)	DFU	CKLG PvW AHL	DKL AR Cooper	9
Pros further	CS	Burns			2
Different	CM				1
Total	4	2	3	4	13

and that is the study of Palacios-Huerta and Volij (2006). On the other hand, both Burns (1985) and Carpenter and Seki (2005) find that professionals and students have different behavior, but the difference goes the other way; that is, students are closer to the theoretical prediction than professionals. In both of these cases, it would seem that the sources of the difference are elements of the professional's work environment or preferences which leads to behavior different from that of the students. As for Cadsby and Maynes (1998a), although neither professionals nor students are closer to the theory, the behavior is so different that it needs to be categorized differently form the other studies where neither professionals nor students are closer to the theory in a way that changes how one thinks of the theory's performance. That is, even though both groups are at a similar distance from the theory, one reaches qualitatively very different conclusions by studying these two groups.

One issue that such a review brings to light is the difficulty of having comparable incentives when comparing students and professionals. For instance, should the average payoffs be the average opportunity cost of time of professionals, or that of students, or should each group receive different incentives? The most thorough approach in that regard might be that adopted in Cooper et al. (1999), where professionals are paid, on average, their opportunity cost of time, while students are divided into one group that faces those same incentives while another group receive payoffs equal to the opportunity cost of time for students.

Another difficulty is that to the extent that differences between professionals and students are observed, although it is tempting to attribute those differences to the experience of professionals, one cannot ignore that the two samples may vary along many other dimensions: gender, race, age, socioeconomic background, and so on. These factors could cause differences which have nothing to do with the fact that one group is composed of professionals and the other is not. In that respect, the most careful paper is Cooper (2006), in which he tries to disentangle the role of these various factors.

It also raises the question, What is the group of interest? Who are the agents that are supposed to be represented in those models being tested? Are Japanese

fishermen organized in cooperative closer to the group of interest than students? What about soccer players? In that last case, for instance, although it is true that soccer players have frequent exposures to situations where randomization is key, they also have more brain lesions than people who do not play soccer (Autti et al., 1997)?

In my view, this survey of the studies that allow us to compare professionals and students, although it does indicate that there are situations where focusing on students is too narrow, does not give overwhelming evidence that conclusions reached by using the standard experimental subject pool cannot generalize to professionals. Studying professionals can prove very insightful in ways that studying undergraduates is not. For instance, the reaction of Chinese managers to context was informative (Cooper et al., 1999), and the fact that professional managers use different messages to get employees out of "bad" equilibria faster (Cooper, 2006) was also an opportunity to learn something that could not have been learned with students. Nonetheless, overall much of the big picture seems the same whether one looks at professionals or students in laboratory experiments testing economic models.

NOTES

I am grateful to Chloe Tergiman for re-creating the tables, and I would like to acknowledge Emanuel Vespa for helping with research. I also wish to thank David Cooper for comments and all those who have sent me papers that I had missed in earlier drafts. I gratefully acknowledge the support from of NSF via grants SES-0519045, SES-0721111, and SES-0924780 as well as support from the Center for Experimental Social Science (CESS), the C.V. Starr Center.

1. Ball and Cech (1996) also reviewed studies comparing students and professionals. However, my review excludes some of the papers they include because they do not fit the guidelines I impose, but more importantly I also cover papers they did not.

2. Studies comparing professionals to students were discarded if their student group was mostly composed of graduate students. Anderson and Sunder (1995) used MBA students. Hong and Plott (1982) used graduate students in engineering, business, and law.

3. Riker (1967, p. 58) was excluded even though it has two groups, one with more experience than the other, because it is unclear who composes that second group. All that is known is that they are male evening students in the University of Rochester graduate "college of business ranging in age from 25 to 55 and in occupation from management trainee and mailman to department head at the largest local industrial plant."
Banks et al. (1994) studied refinements in signaling games and compared students to members of the technical staff at the Jet Propulsion Laboratory. The employees of the Jet Propulsion Laboratory must have a high level of technical and mathematical expertise, but since those attributes do not seem to make them professionals in signaling, this study was not included either.

4. Some studies were excluded on the basis that there were not enough details on experimental procedures or because the treatments with professionals had variations with

respect to the treatments with students. These include Mestelman and Feeny (1988), Smith et al. (1988), and King et al. (1992).
The only variations in treatment between students and professionals that were allowed were those having to do with incentives and the number of rounds per sessions.

5. However, I have included a couple of experiments with incentives that are not typical by experimental economics standards (the Burns (1985) study as well as the Dejong et al. (1988)). Those study are included mainly because they are some of the oldest. However, the unusual incentives should be kept in mind when considering the results. The requirement of neutral language eliminated studies such as Alatas et al. (2006).

6. Based on this last criterion, Montmarquette et al. (2004) was not included although it meets all of the other criterions.

7. It is not always obvious that one cares only about the behavior of professionals. For one thing, there are situations for which there are no obvious professionals (e.g., voluntary contributions to charity). Second, sometimes they are not the only groups of interest (e.g., investment behavior by small individual investors). Third, there are situations that people face only once in their life, and thus they are by definition inexperienced when they play the relevant game (e.g., medical students going through the residency match).

8. Most of the substance here is from Shadish et al. (2002).

9. This could be the result, for instance, of professionals using, in the way they think about what to do, elements which matter in their professional environment even though they are not in the particular environment of the experiment.

10. A more detailed analysis of the students experiment is available in Fouraker et al. (1962).

11. In the ultimatum game, a proposer offers a division of an amount of money and the responder can accept, in which case both players receive the proposed split, or reject, in which case they both receive nothing.

12. Although I will simply base my analysis on the authors own account of their data, Wooders (2008) reaches different conclusions from reanalyzing the Palacios-Huerta and Volij (2006) data.

REFERENCES

Abbink, Klaus and Bettina Rockenbach. 2006. Option Pricing by Students and Professional Traders: A Behavioral Investigation. *Managerial and Decision Economics* **27**:497–510.

Alatas, Vivi, Lisa Cameron, Ananish Chaudhuri, Nisvan Erkal, and Lata Gangadharan. 2006. Subject Pool Effects in a Corruption Experiment: A Comparison of Indonesian Public Servants and Indonesian Students. The University of Melbourne Research Paper Number 975. URL: http://www.economics.unimelb.edu.au/SITE/research/workingpapers/wp06/975.pdf.

Alevy, Jonathan E., Michael S. Haigh, and John A. List. 2009. Information Cascades: Evidence from a Field Experiment with Financial Market Professionals. *Journal of Finance* **62**(1):151–180.

Andersen, Steffen, Glenn Harrison, Morten Lau, and Elisabet Rutstrom. 2010. Preference Heterogeneity in Experiments: Comparing the Field and Lab. *Journal of Economic Behavior and Organization*. *Journal of Economic Behavior & Organization* **73**(2):209–224.

Anderson, Lisa R. and Charles A. Holt. 1997. Information Cascades in the Laboratory. *American Economic Review* **87**(5):847–862.

Anderson, Matthew J. and Shyam Sunder. 1995. Professional Traders as Intuitive Bayesians. *Organizational Behavior and Human Decision Processes* **64**(2):185–202.

Autti, T., L. Sipilä, H. Autti, and O. Salonen. 1997. Brain Lesions in Players of Contact Sports. *The Lancet* **349**(19):1144.

Ball, Sheryl B. and Paula-Ann Cech. 1996. Subject Pool Choice and Treatment Effects in Economic Laboratory Research. In *Research in Experimental Economics*, Volume 6. Greenwich, CT: JAI Press, pp. 239–292.

Banks, Jeffrey, Colin Camerer, and David Porter. 1994. An Experimental Analysis of Nash Refinements In Signaling Games. *Games and Economic Behavior* **6**:1–31.

Berg, Joyce, John Dickaut, and Kevin McCabe. 1995. Trust, Reciprocity and Social History. *Games and Economic Behavior* **10**:122–142.

Burns, Penny. 1985. Experience and Decision Making: A Comparison of Students and Businessmen in a Simulated Progressive Auction. In *Research in Experimental Economics*, Volume 3. Greenwich, CT: JAI Press, pp. 139–153.

Cadsby, Charles Bram and Elizabeth Maynes. 1998a. Choosing Between a Socially Efficient and Free-Riding Equilibrium: Nurses Versus Economics and Business Students. *Journal of Economic Behavior and Organization* **37**:183–192.

Cadsby, Charles Bram and Elizabeth Maynes. 1998b. Gender and Free Riding in a Threshold Public Goods Game: Experimental Evidence. *Journal of Economic Behavior and Organization* **34**:603–620.

Camerer, Colin F. 1995. Individual Decision Making. In *Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press, 587–703.

Campbell, Donald T. 1957. Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin* **54**(4):297–312.

Carpenter, Jeffrey and Erika Seki. 2005. Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyoma Bay. IZA DP No. 1697.

Casari, Marco, John C. Ham, and John H. Kagel. 2007. Selection Bias, Demographic Effects and Ability Effects in Common Value Auction Experiments. *American Economic Review* **97**(4):1278–1304.

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Rand McNally.

Cooper, David. 2006. Are Experienced Managers Experts at Overcoming Coordination Failure? *The B.E. Journal of Advances in Economic Analysis and Policy* **6**(2).

Cooper, David J., John H. Kagel, Wei Lo, and Qing Liang Gu. 1999. Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers. *American Economic Review* **89**:781–804.

Cooper, David J., Susan Garvin, and John H. Kagel. 1997. Signalling and Adaptive Learning in an Entry Limit Pricing Game. *RAND Journal of Economics* **28**:662–683.

Croson, Rachel and Melanie Marks. 2001. The Effect of Recommended Contributions in the Voluntary Provision of Public Goods. *Economic Inquiry* **39**(2):238–249.

Croson, Rachel T. A. and Melanie Beth Marks. 2000. Step Returns in Threshold Public Goods: A Meta- and Experimental Analysis. *Experimental Economics* **2**:239–259.

Davis, Douglas D. and Charles A. Holt. 1993. *Experimental Economics*. Princeton, NJ: Princeton University Press.

Dejong, Douglas V., Robert Forsythe, and Wilfred C. Uecker. 1988. A Note on the Use of Businessmen as Subjects in Sealed Offer Markets. *Journal of Economic Behavior and Organization* **9**:87–100.

Dyer, Douglas and John H. Kagel. 1996. Bidding in Common Value Auctions: How the Commercial Construction Industry Corrects for the Winner's Curse. *Management Science* **42**:1463–1475.

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):108–115.

Fehr, Ernst and John A. List. 2004. The Hidden Costs and Returns of Incentives—Trust and Trustworthiness Among CEOs. *Journal of the European Economic Association* **2**(5):743–771.

Fouraker, Lawrence E., Sidney Siegel, and D. L. Harnett. 1962. An Experimental Disposition of Alternative Bilateral Monopoly Models under Conditions of Price Leadership. *Operations Research* **10**(1):41–50.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. Performance in Competitive Environments: Gender Differences. *Quarterly Journal of Economics* **CXVIII**:1049–1074.

Gode, Dhananjay K. and Shyam Sunder. 1993. Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. *Journal of Political Economy* **101**(1):119–137.

Harbaugh, William T., Kate Krause, and Timothy R. Berry. 2001. GARP for Kids: On the Development of Rational Choice Behavior. *American Economic Review*: **91**(5):1539–1545.

Holt, Charles A. 1995. Industrial Organization: A Survey of Laboratory Research. In *Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press, 349–444.

Hong, J. T. and Charles R. Plott. 1982. Rate Filing Policies for Inland Water Transportation: An Experimental Approach. *The Bell Journal of Economics* **13**:1–19.

John H. Kagel, Ray C. Battalio, R.C., and Leonard Green. 1995. *Economic Choice Theory: An Experimental Analysis of Animal Behavior.* New York: Cambridge University Press.

Kagel, John H. 1987. Economics According to the Rats (and Pigeons Too); What Have We Learned and What Can We Hope to Learn? In *Laboratory Experimentation in Economics: Six Points of View*, ed. Alvin E. Roth. New York: Cambridge University Press, pp. 155–192.

Kagel, John H. and Dan Levin. 2002. *Common Value Auctions and the Winner's Curse.* Princeton University Press.

King, Ronald R., Vernon L. Smith, Arlington W. Williams, and Mark Van Boening. 1993. The Robustness of Bubbles and Crashes in Experimental Stock Markets. In *Evolutionary Dynamics and Nonlinear Economics—A Transdisciplinary Dialogue.* Oxford, England: Oxford University Press, 183–200.

Ledyard, John O. 1995. Public Goods. In *Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press, pp. 111–194.

Marks, Melanie and Rachel Croson. 1998. Alternative Rebate Rules in the Provision of a Threshold Public Good: An Experimental Investigation. *Journal of Public Economics* **67**:195–220.

Mestelman, Stuart and David Feeny. 1988. Does Ideology Matter?: Anecdotal Experimental Evidence on the Voluntary Provision of Public Goods. *Public Choice* **57**:281–286.

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):1379–1389.

O'Neil, Barry. 1987. Nonmetric Test of the Minimax Theory of Two-Person Zerosum Games. *Proceedings of the National Academy of Sciences* **84**:2106–2109.

Palacios-Huerta, Ignacio and Oscar Volij. 2006. Experientia Docet: Professionals Play Minimax in Laboratory Experiments. Working paper. URL: http://www.najecon.org/naj/cache/12224700000000l050.pdf.

Potamites, E. and B. Zhang. 2012. Measuring Ambiguity Attitudes: A Field Experiment among Small-Scale Stock Investors in China. *Review of Economic Design* **16**(2-3):193–213.

Potters, Jan and Frans van Winden. 2000. Professionals and Students in a Lobbying Experiment: Professional Rules of Conduct and Subject Surrogacy. *Journal of Economic Behavior and Organization* **43**:499–522.

Riker, William H. 1967. Experimental Verification of Two Theories About *n*-Person Games. In *Mathematical Applications in Political Science III*, ed. Joseph L. Bernd. Charlottesville, VA: University of Virginia Press, pp. 52–66.

Roth, Alvin E. 1995. Bargaining Experiments. In *Handbook of Experimental Economics*, eds. John H. Kagel and Alvin E. Roth. Princeton, NJ: Princeton University Press, pp. 253–348.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Siegel, Sidney and D. L. Harnett. 1964. Bargaining Behavior: A Comparison Between Mature Industrial Personnel and College Students. *Operations Research* **12**(2):334–343.

Smith, Vernon L., G. L. Suchanek, and A. W. Williams. 1988. Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets. *Econometrica* **56**:1119–1151.

Van Huyck, John, Raymond C. Battalio, and Richard Beil. 1990. Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. *American Economic Review* **80**(1):234–248.

Wooders, John. 2008. Does Experience Teach? Professionals and Minimax Play in the Lab. Working paper. URL: http://www.u.arizona.edu/~jwooders/Reexamination 2006 2009 202008.pdf.

CHAPTER 18

THE EXTERNAL VALIDITY OF LABORATORY EXPERIMENTS: THE MISLEADING EMPHASIS ON QUANTITATIVE EFFECTS

JUDD B. KESSLER
AND LISE VESTERLUND

Introduction

Laboratory experiments are used to address a wide variety of questions within economics, including whether behavior is consistent with the predictions and assumptions of theory and how various mechanisms and institutions affect the behavior of economic agents (see Roth (1987) for an overview). The experimental laboratory has become an integral part of the field of economics and a productive dialog now exists between theory, laboratory experiments, and field studies.

Recently, however, a set of papers by Levitt and List (2006, 2007a, 2007b, 2008) has questioned what we can learn from laboratory studies. At the center of their critique is the statement that "the critical assumption underlying the interpretation of data from lab experiments is that the insights gained can be extrapolated to the world beyond" (Levitt and List, 2007a, p. 153) and the subsequent argument that there are "many reasons to suspect that these laboratory findings might fail to generalize to real markets" (Levitt and List, 2008, p. 909), suggesting that the critical assumption about extrapolation may not hold. Specifically, the authors highlight five factors that differ between the laboratory and "the world beyond."[1] They recognize that external validity is also a concern for field experiments and for naturally occurring data; however, their statement that "field experiments avoid many of the important obstacles to generalizability faced by lab experiments" (Levitt and List, 2008, p. 910) has caused many to interpret their papers as an attempt to discredit laboratory experiments and to rank field experiments as a superior methodology.

The papers by Levitt and List have caused quite a stir both inside and outside of the field of experimental economics. A common response in defense of laboratory experiments has been to counter attack field experiments, arguing that field experiments suffer from the same charges levied at laboratory experiments, namely a lack of external validity.[2]

In his reply to Levitt and List, Camerer (2015) moves beyond the generalizability of field experiments and systematically addresses the five factors that Levitt and List (2007a) argue reduce the generalizability of laboratory studies. Camerer (2015) notes that the features of the lab that differ from the field make less of a difference on behavior than Levitt and List (2007a) suggest. He comments that when concerns about one of these factors arise, lab studies can be altered to better mirror an external environment of interest.[3] In addition, he compares the strengths and weaknesses of data collected in the lab and field, arguing that laboratory experiments are more easily replicated whereas field experiments are less obtrusive. However, Camerer also makes the striking argument that considering external validity is distracting for a large class of laboratory studies. While he states that external validity is crucial for studies that aim to inform policy, he argues that it is not necessary for studies aiming to understand general principles. Referring to the former as the policy view and the latter as the science view, Camerer (2015) argues that since experiments conducted under the science view do not aim to forecast behavior in a particular external target setting, we should not care whether these laboratory results generalize to the field.

The papers by Levitt and List and the reply by Camerer (2015) contribute to what many view as an overdue debate on the contribution of laboratory experiments to economics. Unfortunately, much of the debate has been aimed at a straw-man version of external validity. While the debate has centered on the extent to which the *quantitative* results are externally valid, we will argue that for most laboratory studies it is only relevant to ask whether the *qualitative* results are externally

valid. A quantitative result refers to the precise magnitude or parameter estimate of an effect, and a qualitative result refers to the direction or the sign of the estimated effect.[4] Interestingly, among the authors on both sides of the debate, there is much less (and possibly no) disagreement on the extent to which the qualitative results of a laboratory study are externally valid.

In the section entitled "Quantitative Versus Qualitative External Validity," we explain why for most laboratory studies it is only relevant whether the *qualitative* or directional results of the study are externally valid. In the section entitled "Do Laboratory Studies Promise Generalizability," we argue that laboratory studies are conducted to identify general principles of behavior and therefore promise to generalize. In the section entitled "Do Laboratory Results Inform Us About the World Outside the Lab," we examine whether laboratory experiments live up to this promise. We discuss the extent to which qualitative results persist outside of the lab, and how we should respond when they do not. We will avoid the debate on whether the concerns about external validity are more or less warranted in laboratory or field environments. We do not see this debate as being productive as it presupposes that the methodologies are in competition. We conclude the paper by arguing, as many others do, that the lab and field methodologies are highly complementary and that both provide important insights to our understanding of economics.[5]

Quantitative Versus Qualitative External Validity

In the debate about whether laboratory studies are "generalizable" or "externally valid," these terms are often not explicitly defined. Indeed, formal definitions of external validity vary substantially. Some definitions of external validity simply require that the *qualitative* relationship between two variables hold across similar environments. For example, Guala (2002, p. 262) states that "an experimental result is internally valid, if the experimenter attributes the production of an effect B to the factor . . . A, and A really is the . . . cause of B in the experimental set-up E . . . [The experimental result] is externally valid . . . if A causes B not only in E, but also in a set of other circumstances of interest F, G, H, etc." More demanding definitions of external validity additionally require that the *quantitative* effect of a one-unit change in A on B identified in one set-up hold in other comparable environments.

Levitt and List (2007a) describe concerns about laboratory experiments meeting the higher standard. In their conclusion, they accept that laboratory experiments meet the first definition of external validity, and they argue that (emphasis added) "lab experiments that focus on *qualitative* insights can provide a crucial first understanding and suggest underlying mechanisms that might be at work when certain data patterns are observed" (pp. 170–171).[6] However, they argue that

laboratory experiments fail to meet the higher standard, questioning whether (emphasis added) "the experimental findings are *equally* descriptive of the world at large" (p. 158). More directly, Levitt and List (2007b, p. 351) write "even for those experiments that are affected by our concerns, it is likely that the qualitative findings of the lab are generalizable, even when the quantitative magnitudes are not." In responding to Levitt and List, the subsequent debate has centered on the extent to which *quantitative* lab findings are externally valid.

This focus on quantitative external validity is misplaced for many (if not most) experimental studies, however, as the emphasis in laboratory studies is to identify the direction rather than the precise magnitude of an effect. Indeed, the nonparametric statistical methods commonly used to infer significance rely solely on qualitative differences. Few experimental economists would argue that the precise magnitude of the difference between two laboratory treatments is indicative of the precise magnitude that one would expect to see in the field or even in other laboratory studies in which important characteristics of the environment have changed.[7] For example, the revenue difference between an English auction and a first-price sealed-bid auction in the lab is not thought to be indicative of the quantitative difference that one would find between any other set of English and first-price sealed-bid auctions. Similarly, despite the clear objective of finding externally valid results, the experiments that tested various designs of FCC spectrum auctions were not aiming to identify magnitudes that would generalize to the field. Instead, they were run with the expectation that the directional effects and general principles of behavior identified in lab auctions also would be present in the field (e.g., Ledyard et al., 1997; Plott, 1997).

The emphasis on qualitative results is in part explained by the fact that all theoretical and empirical models require simplifying assumptions. In constructing these models, we eliminate any factors that we think are not central. A consequence of abstracting away from environments of interest is that we likely fail to capture the precise magnitude of the effect we expect to see in those environments.[8]

Since most experimental studies focus on directional effects, the debate about external validity should center on qualitative rather than quantitative predictions. Falk and Heckman (2009) introduce a framework that we can use to conceptualize the difference between the two types of external validity. Considering a relationship between an outcome Y and a number of variables as defined by $Y = f(X_1, X_2, \ldots, X_N)$, which they refer to as an all-causes model since it "captures all possible causes of Y," they note that the causal effect of X_1 on Y is the effect of varying X_1 holding fixed $X^* = (X_2, \ldots, X_N)$.[9] Following Levitt and List, Falk and Heckman also focus on the conditions under which the quantitative findings of the laboratory are externally valid. They show that the substantial requirements for quantitative external validity are that f is separable in X_1 and Y is linear in X_1. Note, however, that the requirements securing external validity of the qualitative effects are weaker. For the qualitative results to be externally valid, we simply require Y to be monotonic in X_1 and for changes in X^* to not reverse the relationship of X_1 on Y.

In the all-causes model, the concerns about external validity raised by Levitt and List are concerns that in the laboratory the magnitude of X_1 and the level at which X^* is held fixed do not correspond to environments outside of the lab. If, in contrast to the current debate, the concern is whether the qualitative effect generalizes, then the differences between the laboratory and the field are only relevant if they are thought to reverse the relationship of X_1 on Y.

Take, for example, the winner's curse. Early experimental demonstrations of the winner's curse, using student subjects, found that increasing the number of bidders increased seller revenue while providing public information about the value of the item for auction decreased seller revenue (Kagel and Levin, 1986). Since experience may influence a bidder's understanding of the incomplete information problem at the core of the winner's curse, the quantitative effect of increasing the number of bidders or the effect of increasing public information may differ across subject pools. However, independent of the subject pool, we expect that increasing the number of bidders will increase the number of individuals who fail to understand the winner's curse, thus increasing the seller's revenue. And we expect that providing public information will mitigate the effect of incomplete information on the bids of anyone who had previously failed to recognize the winner's curse, thus decreasing the seller's revenue. The magnitude of these comparative statics could be different between students and, say, oil company executives, but we expect the qualitative results to be the same.[10]

Do Laboratory Studies Promise Generalizability?

In the debate about the external validity of laboratory experiments, there has been disagreement about when external validity is important. Levitt and List state that "the critical assumption underlying the interpretation of data from lab experiments is that the insights gained can be extrapolated to the world beyond." Schram (2005, abstract) has a more moderate statement noting: "External validity is relatively more important for experiments searching for empirical regularities than for theory-testing experiments." Camerer (2015) takes this argument one step further and argues that external validity is important for experiments conducted from a policy view, but not for experiments conducted from a scientific view.[11] Camerer (2015) states "since the goal is to understand general principles, whether the 'lab generalizes to the field' . . . is distracting, difficult to know . . . and is no more useful than asking whether 'the field generalizes to the lab'" (p. 3).[12]

While there may be disagreement on whether there is a promise for quantitative results of an experiment to be externally valid, we do not think that there can be much disagreement on the extent to which the qualitative results promise external validity. Since laboratory experiments are meant to uncover general principles of

behavior, it is difficult to see how a concern for external validity is not warranted. Even without a particular external target in mind, the general rules that govern behavior in the experimental environment must apply in other environments with similar characteristics.[13] Surely it is a minority of experimental studies that examine environments that have no counterpart outside of the study and for which we would not expect that the "insights gained can be extrapolated to the world beyond." While laboratory studies may not promise quantitative external validity, they do promise qualitative external validity. The question of interest is whether they live up to this promise.

Do Laboratory Results Inform Us About the World Outside the Lab?

Over the course of the debate, authors have suggested two conditions under which we can extrapolate from the laboratory to other environments of interest. Falk and Heckman (2009) summarize the two conditions: "When the exact question being addressed and the population being studied are mirrored in an experiment, the information from it can be clear and informative. Otherwise, to transport experimental findings to new populations or new environments requires a model." Camerer (2015) also highlights that extrapolation is warranted either when the population and environment examined in the laboratory mirrors an environment of interest or when one uses previous studies to account for differences between the lab and field (implying one has an underlying model in mind). Camerer (2015) states "parallelism does not require that students in a lab setting designed to resemble foreign exchange behave in the same way as professional foreign exchange traders on trading floors. . . . The maintained assumption of parallelism simply asserts that if those differences could be held constant (or controlled for econometrically), behavior in the lab and the trading floor would be the same." Levitt and List (2007b) also stress the value of a model in extrapolating experimental results when they write "even in cases where lab results are believed to have little generalizability, some number [a laboratory estimate] is better than no number, provided the proper theoretical model is used to make inference" (p. 364).[14]

While mirroring a particular environment of interest or using a model for inference are both appealing, it is important to recognize that these conditions are very stringent. It is difficult to envision a laboratory study that fully mirrors the circumstances of the external environment of interest, and it is unrealistic to think that we can find a model that allows us to predict how differences between the lab and the field will interact with any comparative static that we observe in the lab. If these conditions were necessary for external validity, then laboratory studies would provide limited insight about behavior outside the lab. Fortunately, neither of these conditions is necessary for the qualitative results of a lab experiment to extrapolate

to other environments of interest. As noted earlier, the qualitative effects will be externally valid if the observed relationship is monotonic and does not change direction when changing the level of variables seen in the field relative to those in the lab.

In a laboratory experiment, subjects are presented with incentives that are meant to capture the central features of the environment in which the economic decisions are usually made. The experimenter has in mind a model that assumes that the laboratory environment does not differ from a comparable field environment on a dimension that would change the sign of the comparative static.[15] Provided that the experimental model is correct, the qualitative results should generalize. What does that mean in practice? What can we conclude about behavior outside the laboratory when we reject, or when we fail to reject, a directional hypothesis in the laboratory?

Suppose that laboratory results reject the hypothesis that a variable affects behavior in a certain way. To what extent does this finding allow us to draw an inference on the role the manipulated variable will have outside of the laboratory? If in the very controlled laboratory setting we reject our hypothesis, then it is unlikely that the manipulated variable will affect behavior in a more complicated external environment.[16]

What if instead we fail to reject a hypothesis in the lab? Does that imply that the hypothesis is likely to find support in field settings with similar characteristics? Schram (2005, p. 232) argues that: "After a theoretical design, a test [of a new airplane] in a wind tunnel is the stage of laboratory experimentation. If it does not 'crash' in this experiment, the plane is not immediately used for the transport of passengers, however. One will typically conduct further tests in the wind tunnel under extreme circumstances. In addition, further testing including 'real' flights without passengers will be conducted." Thus, finding lab evidence consistent with a theory will typically lead to repeated investigations of the result, and ideally these will be done under various stress-test conditions in the lab and in the field. Absent these stress tests, however, is it reasonable to expect the documented comparative static in the lab to also hold in the field? The answer may depend on the strength of our prior, but identifying a comparative static in the lab certainly increases our posterior belief that the comparative static will be found in the field. Since the lab is thought to investigate general principles of behavior, we expect these principles to hold both inside and outside of the laboratory.

As with any finding, however, caution is needed to generate predictions in different settings. For example, in documenting statistical discrimination against women in the sports card market, List (2004) does not claim that women always will be charged a price that is a specific magnitude greater than that for men, or that women always will be charged a higher price, or that there always will be statistical discrimination, but rather that when there are grounds for statistical discrimination against a particular group, the market is likely to respond in a predictable way. For example, in a study on taxi fare negotiations in Lima, Peru, Castillo et al. (2013) confirm this prediction by showing that statistical

discrimination leads to inferior outcomes for men since they have a greater willingness to pay for taxi rides than do women.

So what should be done if we identify a comparative static in the lab but fail to find evidence of the comparative static outside of the lab?[17] When designing an experiment, the experimenter assumes that the lab setting captures the important characteristics of environments of interest and that the qualitative result will hold outside the lab. Failure to replicate a lab finding in the field may result from the experimenter's model failing to capture central features of the decision environment outside the lab. That is, it is not a question of the laboratory failing to identify general principles of behavior, but rather a question of the laboratory model not capturing the external environment of interest. This is akin to when a result that holds true in a model is not observed in the world. In these cases, we infer that the model has an assumption that does not hold or that the model has abstracted away from something important. Consequently, failure to replicate an experimental finding should cause us to revisit the question at hand, as it may be an indication that the laboratory and field environments were different on a dimension that plays an important role in driving the comparative static results.

For example, theoretical studies by Goeree et al. (2005) and Engers and McManus (2007) along with lab studies by Orzen (2008) and Schram and Onderstal (2009) all demonstrated that all-pay charity auctions generated higher revenue than other fundraising mechanisms, while subsequent field studies contradicted this comparative static. Carpenter et al. (2008) and Onderstal et al. (2013) both found that contributions fell under the all-pay auction. Interestingly, the field studies also demonstrated why this discrepancy may have occurred. While the theory and laboratory experiments had assumed full participation, the field studies found that potential donors will opt out of participating in the all-pay auction. Thus the inconsistencies between the lab and field resulted from an incorrect (and restrictive) assumption of full participation in the auction.[18] By ignoring the importance of participation, the initial laboratory model did not capture an essential feature of the external environment of interest.

When results of a laboratory study are not observed in certain field settings, it is of interest to determine which assumption in the laboratory has failed to hold true. The fact that certain laboratory environments may fail to capture the central features of the decision environment outside the lab is raised in Kessler (2013), which highlights a distinction between differences arising due to methodology (e.g., in laboratory experiments subjects know they are part of an experiment and have volunteered to participate, which need not be the case in field experiments) and differences arising due to the economic environment of interest (e.g. information, incentives, action spaces, and so on) that might vary across settings and can be manipulated by an experimenter.

While it is tempting to conclude (as Levitt and List 2007a do) that inconsistencies between lab and field studies result from methodological differences, care should be given to determine whether instead differences controlled by the

experimenter could be driving the results. Kessler (2013) aims to explain why gift exchange is more commonly seen in laboratory experiments than in field experiments. Using a laboratory experiment, he shows that differences in the relative wealth of the firm and the efficiency of worker effort (choices of the experimenter, not differences due to methodology) contribute significantly to the differences in results between the laboratory studies and the field studies. Another example is the lab and field differences of Dutch and sealed-bid auctions. While laboratory studies by Cox et al. (1982, 1983) find that the revenue in sealed-bid auctions dominates that in Dutch auctions, a field study by Lucking-Reiley (1999) finds the reverse revenue ranking. While these results initially were ascribed to methodological differences between lab and field, a subsequent laboratory study by Katok and Kwasnica (2008) shows that choices in the study designs of the previous experiments can explain the divergent results. Specifically, they note that the clock speed in Lucking-Reiley was much slower than that in Cox et al., and they show that revenue in the Dutch auction is significantly lower than in the sealed-bid auction at fast clock speeds, whereas the reverse holds at slow clock speeds.[19] As the initial study failed to account for the effect of clock speed on the revenue ranking, the model was misspecified and the results seen at fast clock speeds did not generalize to environments with slow clock speeds.

Note that in these examples the failure of results to generalize between laboratory and field was *not* a failure of laboratory methodology but rather evidence that the laboratory experiment and field experiment differed on an important feature of the decision environment. By identifying which features of the decision environment are causing the differential results (and which are not), we hone our model of behavior.[20]

Conclusion

Economic research aims to inform us of how markets work and how economic agents interact. Principles of economic behavior are expected to apply outside of the unique environment in which they are identified. The expectation and promise of economic research is that the uncovered principles of behavior are general and therefore externally valid. However, this promise does not imply that the magnitude of an estimated effect applies generally. In many cases, including many experimental economics studies, the expectation is simply that the qualitative or directional results are generalizable. The simplifying assumptions used to secure internal validity imply that the magnitude of the observed effect will likely differ from the magnitudes in other environments. Interestingly, there appears to be broad agreement that the qualitative results seen in the laboratory are externally valid. To our knowledge, there is no evidence suggesting that the lab–field differences discussed in the ongoing debate reverse directional effects identified in the lab.

In emphasizing the importance of qualitative results, we have ignored the studies that appear to estimate preference parameters in the laboratory. The objective of some of these studies is to derive comparative statics, whereas others emphasize the parameter estimates themselves; and while some of these parameter estimates may be thought to be scale-free and generalizable, others are context dependent and therefore unlikely to generalize.

When authors use preference parameters to generate comparative statics, they often do so with the expectation that the comparative statics, rather than the estimated preference parameters, will generalize. For example, while Andreoni and Vesterlund (2001) estimate male and female demand functions for giving in the laboratory, they solely emphasize the surprising comparative static result that women are less sensitive than men to the price of giving, and it is this comparative static that they subsequently try to extrapolate. They first note that Andreoni et al. (2003) find the same gender difference in price sensitivity when examining how annual giving responds to an individual's marginal tax rate. Then, using data on tipping by Conlin et al. (2003), they find that tipping by men is more sensitive to the cost of tipping than it is for women. Thus, despite generating demand estimates for giving, Andreoni and Vesterlund (2001) do not examine whether the quantitative results generalize, instead they use the qualitative results to predict behavior outside the laboratory. This emphasis on comparative statics is also seen in some studies on individual risk preferences, time preferences, and preferences over payoffs to others, which aim to identify general principles such as loss aversion, probability weighting, present bias, and inequality aversion.

While we may expect the comparative statics derived from preference parameter estimates to generalize, it is questionable whether the estimates themselves will generalize. For example, while lab and field studies on other-regarding preferences help identify the general characteristics of behavior that result from such preferences, they are unlikely to identify the magnitude of such effects across domains. Considering the amount of work professional fundraisers put into soliciting funds, it is clear that other-regarding behavior depends greatly on context. One act of charity by an individual cannot predict all his other charitable acts; instead, each charitable act has specific characteristics. Hoping that an estimated preference for giving can be extrapolated to all other environments is similar to hoping that we can predict a consumer's demand for all goods from an estimate on demand for one good.

Perhaps because other-regarding preferences are so complex, it would be particularly costly to dismiss a research methodology from shedding light on the phenomenon. Indeed, research from both lab and field experiments have played a significant role in improving our understanding of what triggers giving. As noted earlier, field experiments helped us understand behavior in all-pay charity auctions. Lab experiments have played an important role in helping us understand charitable giving by providing a controlled environment that enables us to identify which mechanisms may be driving behavior.

For example, field studies have repeatedly shown that contributions in many settings can be impacted by information about the contributions made by previous donors (see, for example, List and Lucking-Reiley (2002), Croson and Shang (2008), Frey and Meier (2004), and Soetevent (2005)). While these studies demonstrate that individuals respond to the contributions of others, they provide little information on which mechanisms may be driving the result. One hypothesis is that information about the contributions of others may provide guidance when there is uncertainty about the quality of the product provided by the organization (e.g., Vesterlund, 2003; Andreoni, 2006). While one easily can show theoretically that sequential giving can generate an increase in donation, signaling is a difficult behavioral task and it may be questioned whether donors will be able to exploit their ability to signal quality. Unfortunately, the signaling model is not easily tested in the field, as it is hard to isolate changes in charity quality. However, it is not difficult to conduct such a study in the laboratory, and indeed a substantial attraction of the lab is that one can easily contrast competing hypotheses. Potters et al. (2005, 2007) investigate sequential giving both with and without uncertainty about the quality of the public good. They find that sequential contributions increase giving when there is uncertainty about the quality of a public good, but not when the quality of the public good is known. Thus, behavior is consistent with individuals seeing large initial contributions as a signal of the charity's quality. This result corresponds with field evidence that information on past contributions has a greater impact on new donors (Croson and Shang, 2008) and a greater impact when contributing to projects for which the donor has less information on quality (Soetevent, 2005).

Lab and field experiments each add unique insights to our understanding of economic behavior. Discussions aiming to secure a relative ranking of the two methodologies are both unwarranted and unproductive. Instead, methodological discussions should highlight the ways in which laboratory and field experiments are complements. And ideally, those discussions will spark new research that takes advantage of their combined strengths.

NOTES

The authors thank George Lowenstein, Jack Ochs, Alvin Roth, and Tim Salmon for their helpful and thoughtful comments, and we thank Guillaume Fréchette and Andrew Schotter for inviting us to write this chapter.

1. The five factors they discuss are: the level of scrutiny, the lack of anonymity, the context, the stakes, and the population.

2. In an echo of the attacks on laboratory experiments, critics have argued that certain markets studied in the field may differ substantially, and thus provide limited insights about, other markets of interest (not coincidentally, a common example has been the sports card market studied in List (2006)). In addition, proponents of laboratory studies have argued that field experiments also lack internal validity as limitations on control in

the field make it more difficult to identify causal relationships. Finally, some have raised concerns about the difficulty of replicating field experiments.

3. Camerer (2015) writes: "We then consider which typical features of lab experiments might threaten generalizability. Are those features a necessary part of all lab experiments? Except for obtrusive observation in the lab (which is inherent in human subjects protection), the answer is 'No!'. The special features of lab experiments which might limit generalizability can be relaxed if necessary to more closely match particular field settings" (p. 3).[1]

4. We thank George Loewenstein for pointing out that these definitions differ substantially from their common uses, in which qualitative refers to things that cannot be measured quantitatively. We use these definitions as they are commonly used in the debate (see, for example, Levitt and List (2007a)).

5. Roth (2008) notes "Lab and field experiments are complements not only with each other, but also with other kinds of empirical and theoretical work." Falk and Heckman (2009) write "Field data, survey data, and experiments, both lab and field, as well as standard econometric methods, can all improve the state of knowledge in the social sciences." In their Palgrave entry on field experiments, Reiley and List (2007) write "the various empirical approaches should be thought of as strong complements, and combining insights from each of the methodologies will permit economists to develop a deeper understanding of our science." Levitt and List (2007b) point to the complementarities in stating "we believe that the sharp dichotomy sometimes drawn between lab experiments and data generated in natural settings is a false one. . . . Each approach has a different set of strengths and weaknesses, and thus a combination of the two is likely to provide more insight than either in isolation" (p. 364). As discussed below, Kessler (2013) highlights a specific way in which laboratory and field results are complements in the production of knowledge.

6. Levitt and List also note that "lab experiments can suggest underlying mechanisms that might be at work when certain data patterns are observed and provide insights into what can happen in other related settings" (2007b, p. 363).

7. While many field experiments are written up to emphasize the magnitude of an estimated effect, it is presumably not the intention of the authors that the level of this magnitude is expected to generalize to other environments. For example, List and Lucking-Reiley (2002) identify a nearly sixfold increase in contributions when they increase seed money for a fundraising goal from 10% to 67%. Few would expect this result to generalize to a sixfold increase in all other charitable campaigns. Presumably the authors do not report this result in their abstract to suggest that it is quantitatively generalizable, but instead report the result: to demonstrate the strength of the effect, to compare it to the strength of the other results in their paper, and to suggest that it is of economic significance.

8. We would only describe quantitative relationships with our models if all the factors we assumed away were irrelevant for the magnitude of the examined effect.

9. Note that in many experimental studies, X_1 is a binary variable indicating different market mechanisms or institutions.

[1] In some instances, the quotes cited in this chapter were those of an earlier version of the Camerer paper and hence differ slightly from the printed version in this volume.

10. Dyer et al. (1989) find that professionals also are subject to the winner's curse. See Fréchette (2015) for a review of studies comparing the behavior of students and professionals. Out of 13 studies that allow comparison of professionals and students in standard laboratory games, he finds only one example where the behavior by professionals is closer to what is predicted by economic theory.

11. Camerer (2015) states that "If the purpose of an experiment is to supply a policy-relevant answer to a particular external . . . setting, then it is certainly valid to ask about how well the experimental setting resembles the target external setting. But this is rarely the case in experimental economics" (p. 7).

12. The many experimental studies on various FCC auction mechanisms demonstrate that policy makers and practitioners are often deeply interested in qualitative effects.

13. For example, Plott (1982) argues that the markets examined in the lab also are real markets and therefore that the general principles of economics demonstrated in the lab should also hold in other markets.

14. Levitt and List (2007b) argue that a model is required to predict outside of the laboratory: "Our approach to assess the properties of the situation is to explore, both theoretically and empirically, how individual behavior changes across judiciously chosen levels of these factors, as moderated by both the task and the agent type. Until this bridge is built between the lab and the field, any argument concerning behavioral consistency might be considered premature" (p. 363). They also note that the demands on this model are rather substantial, "unless considerable changes are made in the manner in which we conduct lab experiments, our model highlights that the relevant factors will rarely converge across the lab and many field settings . . . what is necessary are a model and a set of empirical estimates to inform us of when and where we should expect lab behavior to be similar to a particular field environment and, alternatively, when we should expect large differences" (p. 364).

15. If a laboratory experiment were expected to generate a result that was specific to the lab (i.e., rather than a result that identified a general principle) such that the sign of the result might change outside the lab, we contend that the experimenter should not have bothered to run the experiment in the first place.

16. For example, Schram (2005) writes: "The bottom line is that there is no reason to believe that a general theory that is rejected in the laboratory would work well in the world outside of the laboratory" (p. 231). Of course this does not mean that the theory being tested is wrong, it just means is that it is not a good approximation of actual behavior.

17. Of course, some studies are conducted in the laboratory precisely because they cannot be conducted in the field. For example, it is difficult to see how a signaling experiment along the lines of Cooper et al. (1997a,b) could be conducted in the field.

18. See also Ivanova-Stenzel and Salmon (2008) for a further illustration that endogenous entry may influence the revenue rankings in auctions. Interestingly, Corazzini et al. (2010) show a similar decrease in participation in the lab when participants in the all-pay public good auction are given heterogeneous endowments.

19. Katok and Kwasnica (2008) note: "The Cox et al. (1982) study used clocks that descended between 0.75% and 2% of their maximum value every second; the Lucking-Reiley (1999) field study used a clock that decreased approximately 5% per day. . . . Since slower auctions impose higher monitoring and opportunity costs on bidders and are generally less exciting, the slow clock may cause the bidders to end the auction early" (p. 346). We thank Tim Salmon for suggesting this example.

20. In fact, if factors like scrutiny, decision context, or characteristics of the actors interact importantly with a comparative static in a way that we do not expect, the fact that we did not expect the interaction means our model is misspecified. In particular, it means that we have left out an important interaction that will be important to include in the model to make predictions. For example, if only women (or only students) were to respond to the incentive of lowered prices, then a model of demand that does not account for gender (or student status) would fail to explain or predict the effect of prices on behavior.

REFERENCES

Andreoni, J. 2006. Leadership Giving in Charitable Fund-Raising. *Journal of Public Economic Theory* **8**:1–22.

Andreoni, J., E. Brown, and I. Rischall. 2003. Charitable Giving by Married Couples: Who Decides and Why Does It Matter? *Journal of Human Resources* **XXXVIII**(1):111–133.

Andreoni, J. and L. Vesterlund. 2001. Which Is the Fair Sex? Gender Differences in Altruism. *Quarterly Journal of Economics* **116**(1):293–312

Camerer, C. 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*, eds. Fréchette, G. R. and A. Schotter. Oxford University Press.

Carpenter, J., J. Holmes, and P. H. Matthews. 2008. Charity Auctions: A Field Experiment. *Economic Journal* **118**:92–113.

Castillo M., R. Petrie, M. Torero, L. Vesterlund. 2013. Gender differences in bargaining outcomes: A field experiment on discrimination. *Journal of Public Economics* **99**:35–48.

Conlin, M., M. Lynn, and T. O'Donoghue. 2003. The Norm of Restaurant Tipping. *Journal of Economic Behavior & Organization* **52**(3):297–321.

Cooper, D. J., S. Garvin, and J. H. Kagel. 1997a. Adaptive Learning vs. Equilibrium Refinements in an Entry Limit Pricing Game. *Economic Journal* **107**(442):553–575.

Cooper, D. J., S. Garvin, and J. H. Kagel. 1997b. Signalling and Adaptive Learning in an Entry Limit Pricing Game. *RAND Journal of Economics* **28**(4):662–683.

Corazzini, L., M. Faravelli, and L. Stanca. 2010. A Prize to Give For: An Experiment on Public Good Funding Mechanisms. *Economic Journal* **120**(547):944–967.

Cox, J. C., B. Roberson, and V. L. Smith. 1982. Theory And Behavior of Single Object Auctions. *Research in Experimental Economics* **2**:1–43.

Cox, J. C., V. L. Smith, and J. M. Walker. 1983. A Test that Discriminates Between Two Models of the Dutch-First Auction Non-Isomorphism. *Journal of Economic Behavior and Organization* **4**:205–219.

Croson, R. and J. Y. Shang. 2008. The Impact of Downward Social Information on Contribution Decisions. *Experimental Economics* **11**(3):221–233.

Dyer, D., J. H. Kagel, and D. Levin. 1989. Resolving Uncertainty About the Number of Bidders in Independent Private-Value Auctions: An Experimental Analysis. *RAND Journal of Economics* **20**(2):268–279.

Engers, M. and B. McManus. 2007. Charity Auctions. *International Economic Review* **48**(3):953–994.

Falk, A. and J. J. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* **326**:535–538.

Fréchette, G. R. 2015. Laboratory Experiments: Professionals Versus Students. In *Handbook of Experimental Economic Methodology*, eds. Fréchette, G. R. and A. Schotter. Oxford University Press.

Frey, B. S. and S. Meier. 2004. Social Comparisons and Pro-Social Behavior: Testing 'Conditional Cooperation' in A Field Experiment. *American Economic Review* **94**:1717–1722.

Goeree, J. K., E. Maasland, S. Onderstal, and J. L. Turner. 2005. How (Not) to Raise Money. *Journal of Political Economy* **113**(4):897–918.

Guala, F. 2002. On the Scope of Experiments in Economics: Comments on Siakantaris. *Cambridge Journal of Economics* **26**(2):261–267.

Ivanova-Stenzel, R. and T. C. Salmon. 2008. Revenue Equivalence Revisited. *Games and Economic Behavior* **64**(1):171–192.

Kagel, J. H. and D. Levin. 1986. The Winner's Curse and Public Information in Common Value Auctions. *American Economic Review* **76**(5):894–920.

Katok, E. and A. Kwasnica. 2008. Time Is Money: The Effect of Clock Speed on Seller's Revenue in Dutch Auctions. *Experimental Economics* **11**:344–357.

Kessler, J. B. 2013. When Will There Be Gift Exchange? Addressing the Lab-Field Debate With Laboratory Gift Exchange Experiments. Working paper.

Ledyard, J. O., D. Porter, and A. Rangel. 1997. Experiments Testing Multiobject Allocation Mechanisms. *Journal of Economics & Management Strategy* **6**:639–675.

Levitt, S. D. and J. A. List. June 2006. What Do Laboratory Experiments Tell Us About The Real World? Working paper.

Levitt, S. D. and J. A. List. 2007a. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives* **21**:153–174.

Levitt, S. D. and J. A. List. 2007b. Viewpoint: On the Generalizability of Lab Behaviour to the Field. *Canadian Journal of Economics* **40**:347–370.

Levitt, S. D. and J. A. List. 2008. Homo Economicus Evolves. *Science* **319**:909–910.

Levitt, S. D. and J. A. List. 2009. Field Experiments in Economics: The Past, the present, and the future. *European Economic Review* **53**:1–18.

List, J. A. 2004. Substitutability, Experience, and the Value Disparity: Evidence from the Marketplace. *Journal of Environmental Economics and Management* **47**(3):486–509.

List, J. A. 2006. The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy* **114**:1.

List, J. A. and D. Lucking-Reiley. 2002. The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign. *Journal of Political Economy* **110**(1):215–233.

Lucking-Reiley, D. 1999. Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet. *American Economic Review* **89**(5):1063–1080.

Onderstal, S., A. J. H. C. Schram, and A. R. Soetevent. 2013. Bidding to give in the field. *Journal of Public Economics* **105**:72–85.

Orzen, H. October 2008. Fundraising Through Competition: Evidence from the Lab. Discussion Paper 2008-11, CeDEx, University of Nottingham.

Plott, C. R. 1982. Industrial Organization Theory and Experimental Economics. *Journal of Economic Literature* **20**(4):1485–1527.

Plott, C. R. 1997. Laboratory Experimental Testbeds: Application to the PCS Auction. *Journal of Economics & Management Strategy* **6**(3):605–638.

Potters, J., M. Sefton, and L. Vesterlund. 2005. After You-Endogenous Sequencing in Voluntary Contribution Games. *Journal of Public Economics* **89**(8):1399–1419.

Potters, J., M. Sefton, and L. Vesterlund. 2007. Leading-by-Example and Signaling in Voluntary Contribution Games: An Experimental Study. *Economic Theory* **33**(1):169–182.

Reiley, D. H. and J. A. List. 2007. *Field Experiments in Economics*. Working paper version of the Palgrave Chapter.

Roth, A. E. 1987. Laboratory Experimentation in Economics, and Its Relation to Economic Theory. In *Scientific Inquiry in Philosophical Perspective*. University Press of America.

Roth, A. E. 2008. *Laboratory vs. Field Experiments: What Can We Learn?* Comments on Levitt-List paper by Al Roth Roundtable discussion, ASSA meetings, Boston, January 8, 2006.

Schram, A. J., Onderstal, S., May 2009. Bidding to Give: An Experimental Comparison of Auctions for Charity. *International Economic Review* **50**(2):431–457.

Schram, A. J. H. C. 2005. Artificiality: The Tension Between Internal and External Validity in Economic Experiments. *Journal of Economic Methodology* **12**(2):225–237.

Soetevent, A. R. 2005. Anonymity in Giving in a Natural Context—A Field Experiment in 30 Churches. *Journal of Public Economics* **89**(11-12):2301–2323.

Vesterlund, L. 2003. The Informational Value of Sequential Fundraising. *Journal of Public Economics* **87**:627–657.

CHAPTER 19

THE LAB AND THE FIELD: EMPIRICAL AND EXPERIMENTAL ECONOMICS

DAVID REILEY

Introduction

I am grateful for this opportunity to reflect on the purpose of experimentation in economic science. I have been asked to comment on two articles, by Kagel and by Harrison, Lau, and Rutstrom. These articles express a sentiment, shared with a number of other laboratory experimentalists, emphasizing "control" as one of the most important aspects of experimental economics. By contrast, when I think of experiments, I think about a source of exogenous variation in data.

My difference in perspective has something to do with my background and training. As a graduate student, I trained as an empirical economist. One of the most important achievements of applied econometrics in recent decades is a deep understanding of the difficulty of causal inference: using observational data to infer causality rather than mere correlation. Empirical labor economists have been exploiting natural experiments for this purpose for several decades. When I learned about experimental economics circa 1995, I enthusiastically joined the experimental fold, seeing experiments as an opportunity to create exogenous variation in situations where we weren't lucky enough for Nature to have provided us with a natural experiment. Over the past 15 years, the profession has accumulated hundreds of

examples of field experiments, a majority of them conducted by people who, like me, trained in empirical (observational) economics.

By contrast, I observe that the majority of laboratory experiments have been conducted by scholars who trained in economic theory, laboratory methods, or both. We field experimentalists are indebted to these laboratory experiments for paving the way and for demonstrating to us that economics is not merely an observational science (as had been assumed for many decades) but also an experimental one.

Field experimentalists are now taking this important message to other empirical economists. Our most important dialogue is with traditional empiricists, demonstrating to them the feasibility of experiments for answering important questions of measurement and causal inference. Which auction format would raise more revenue for the types of collectibles sold on eBay? Which fundraising strategy would raise more revenue for a public-radio station? Does advertising cause incremental sales? What is the price elasticity of demand for soda? Does an amusement park truly set the monopoly price for soda within the park? What is the magnitude of racial discrimination against African-American job applicants? What are the productivity impacts of incentive-compensation schemes for salespeople? These are questions that previously would have been answered only by analysis of observational data, which is usually woefully inadequate. (How often does the U.S. Forest Service change its timber-auction format? And on the rare instance when they do, how do we know that the auction format is uncorrelated with the quality of the timber being auctioned?) In my view, field experiments are about the creative use of experiments in situations where economists previously assumed their science was purely observational.

"Lack of control" was one of the earliest complaints I heard about field experiments. And I agree: certainly a field experiment provides less control than a laboratory experiment. In auction field experiments, we do not get to control the bidders' values for the good, nor do we even get to observe them directly. Thus, the laboratory provides the valuable ability not only to manipulate the environment, but even to observe it. Field experimentalists know that we give up valuable control in order to get increased realism, such as when we have subjects bid for real goods instead of induced values.

Because of my interest in engaging in dialogue with empirical economists, I do not always see "control" as an unambiguously good thing. Here are three reasons illustrating why control might be undesirable. First, many economic decisions take more time to reach than the typical time limits in a laboratory session. Second, despite careful laboratory protocols to prevent communication, people often do talk to others when making real-world decisions. Third, we do not know whether economic theory is correct in its primitive assumptions, such as modeling charities as public goods, or bidder values to be privately known in auctions. If it turns out that such features matter for economic decision-making, we will only know the truth by going outside the traditional laboratory setting.

John Kagel: "Laboratory Experiments"

John Kagel's paper provides a great summary of the relationship of the laboratory to the field on two important topics: the winner's curse and gift exchange. On the winner's curse, Kagel draws lessons from experiments comparing students with professional construction bidders, and he also makes comparisons to available field data. On gift exchange, Kagel compares laboratory experiments with field experiments, where the field experiments differ from the lab in offering payment for work done outside the laboratory.

Regarding the winner's curse, Kagel reviews his own work, dating back to Dyer et al. (1989) in which he and his coauthors recruited professional construction bidders to participate in abstract laboratory experiments involving common-value auctions. This extremely innovative work was the first example I am aware of to make a serious effort to examine the importance of subject pools in laboratory experiments.

Because of their professional experience estimating anticipating construction costs and bidding based on their estimates, we might expect these subjects to behave very differently from students concerning the winner's curse. However, the experiments show that, if anything, the professional bidders bid even more aggressively than students, falling even more prey to the winner's curse overall.

With hindsight, we can see that the abstract task of the laboratory is a rather different task than what the professional construction managers face in practice, as documented in Dyer and Kagel (1996). In the real world, a number of other concrete factors affect managers, but are missing from the abstract environment of the lab. First, managers have a feel for their own capacity for error in estimating the cost of a real construction job, but they have no experience dealing with the abstract "signals" in a laboratory common-value experiment. Second, in real construction auctions it turns out that bidders who significantly underestimate costs are often able to withdraw their bids without penalty after the auction results are realized, but such protection is not offered to bidders in the abstract laboratory setting. Third, the common-value component in construction auctions appears (upon investigation) to be much less important than theorists might have assumed; instead, privately known differences in firms' unused capacity play a much larger role than the uncertainty of their estimates of uncertain job cost.

As Kagel notes, learning about the importance of these institutional features of the construction industry was a major benefit to the research program comparing professionals to students in the lab. This is a key role of both experimental and empirical economics. Theorists often are unable to say exactly how well their models apply to real-world situations. For auctions in particular, it is very difficult to know how much private uncertainty in values exists in a given auction market; hence it is difficult to know whether private-value or common-value models are

more applicable. Theorists have commonly cited construction as a clear example of a market with a significant common-value component, but upon further empirical and experimental analysis this becomes much less clear. I wish to highlight the importance of collecting real facts about markets and behavior, as well as cite the work by Kagel and coauthors as a prime example of a research program that does just that.

I also want to bring attention to the important role of context in laboratory experiments. Construction bidders showed no better mastery of the winner's curse than students did in an abstract experiment. The bidding skills the professionals have in the field, where we believe they know something about their ability to underestimate costs, do not necessarily transfer to a task where they are asked to understand the variance of an abstract probability distribution.

That is, context matters quite a bit, in the sense that people find it much easier to make good decisions when the setting is more concrete and more familiar. This reinforces results from psychology, notably the Wason (1966) task, where concrete details (such as "Are all the females old enough to drink?") enable subjects to understand a logic puzzle much more readily than in a completely abstract setting (such as "Do all the even-numbered cards have a consonant on the back?"). In Levitt et al. (2010), my coauthors and I show that professional poker players and professional soccer players, despite their experience in games where mixing of strategies is important to success, show just as much deviation from mixed-strategy Nash equilibrium as do student subjects. Because of our belief in the importance of context, we set out to replicate the surprising results of Palacios-Huerta and Volij (2008) that professional soccer players played exactly the Nash proportions in abstract mixed-strategy matrix games. We were able to show that Nash equilibrium is a decent predictor of the average mixing proportions, but unlike in Palacios-Huerta and Volij (2008), we find that professionals have just as much variance away from Nash equilibrium as student subjects do. As Kagel notes, "one should not automatically expect economic agents who are fine tuned to the field settings they typically operate in to be able to adjust to experimental environments that deprive them of their typical contextual cues."

This is one reason why I have been so excited to promote field experiments during my career. By doing field experiments in real markets for real goods and services, we can ensure that subjects have a familiar context and are more likely to behave naturally. Learning about any new economic environment typically takes place only slowly. I can easily find numerous examples of slow learning from my own life, each of which cost me thousands of dollars' worth of utility. I bought three houses using my real estate agent's recommended mortgage lender before it occurred to me that I should shop around for the lender with the best interest rate. I underestimated the benefits I would have received from a prenuptial agreement in my second marriage, in part because my first divorce had been so amicable. When I changed employers, I bought a bare-bones health insurance plan, not realizing that my preferred doctor would subsequently drop that plan because

it was not financially generous enough to him. Of course, one hypothesis is that I am unrepresentatively slow to understand new economic decision problems, but I believe that the experimental data on the importance of context allow us to reject that hypothesis.

Regarding gift exchange, Kagel's second example about interplay between the laboratory and the field, I believe that the laboratory experiments in this area have been extremely interesting and important in documenting reciprocity at odds with standard noncooperative game theory. I also believe that field experiments have been very important for probing the limits of the laboratory results, finding that positively reciprocal behavior can diminish over timescales longer than the typical laboratory session. Kagel's review has convinced me that this issue is far from settled, and we need much more research in this area, in a greater variety of field settings.

Kagel also gives a great summary of the advantages of laboratory experiments, which I think bears repeating here. Much less is observable in a laboratory experiment than in a field experiment: the cost of effort, the value of increased of effort to the employer, and the beliefs that employees have about the game that they are playing. There are also more options available to the experimenter, in the sense that once subjects are in the laboratory, we can ask them if they would like to do "work" at wages that would be unacceptably low in a field experiment, either for legal reasons (minimum wage) or for ethical reasons (IRB) or both. I particularly like Kagel's point that we need to think carefully about the baseline level of wages in future work, because a higher-than-normal wage may already be eliciting positive reciprocity in the control group of a field experiment, even before we implement a treatment that surprises some subjects with even higher wages.

I also like Kagel's observation that, unlike in the laboratory, workers in field settings typically can respond to wage offers along multiple dimensions (quality, quantity, etc.). Kagel assumes that this is to be a disadvantage of field experiments relative to the lab, but it can also be seen as an advantage. If workers in the real world can respond along multiple dimensions, then many things can happen. For example, workers might have some confusion and contribute positively reciprocal actions along a dimension that turns out to have little value for the firm, thereby diluting the positive effect of paying "efficiency wages" relative to what we might estimate in the laboratory setting. It is absolutely an advantage of the laboratory to be able to create simplified, abstract environments that help us isolate individual behavioral effects. However, the richness of the real world demands that we complement our laboratory experiments with field experiments that allow us to estimate causal effects even in the messy real world. Laboratory experiments allow us to isolate precise causal effects on behavior in the simplified models assumed by theory, while field experiments allow us to examine whether the theoretical models have captured the most important aspects of a messy reality.

I agree with Kagel that laboratory experiments have documented the importance of reciprocity in economic transactions and that this provides

valuable information to theorists writing macroeconomic models involving sticky-downward wages. As an empirical economist, I would like to go even further and begin to *measure* the extent to which we should expect wages to be sticky in the real-world economy. The field experiment of Gneezy and List (2006) has made a first attempt in that direction. Knowing what fraction of subjects engage in positive or negative reciprocity in gift-exchange laboratory experiments does not get us very far toward understanding how far labor demand can decrease in the real world before wages must begin to fall.

Kagel is pessimistic about prospects for greater verisimilitude, asserting that "the very structure of these experiments, both laboratory and field studies, cannot approach the target environment of ongoing labor relations within firms" and that being fully faithful to the real-world environment is "an impossibility under any circumstances." (Harrison et al. (2015) similarly express in their paper, in a different context, "There is simply no way to run a naturally-occurring field experiment in this case.") However, I am more optimistic. Bandiera et al. (2005) have already accomplished some very nice field experiments on employee compensation within ongoing firms. More generally, we have been making big, creative advances in the use of field experiments in recent years, conducting real-world experimental measurements in many areas that would have been assumed impossible just 15 years ago: auctions, charitable giving, incentive compensation, and economic-development policy are a few notable examples. I wouldn't be surprised if a clever researcher manages to design a field experiment that gets us much closer to measuring real-world labor demand. I look forward to additional field experiments that may get us closer to an answer to this messy, but important, question.

Harrison, Lau, and Rutstrom: "Theory, Experimental Design and Econometrics are Complementary"

Let me begin by noting, for the record, my (long-standing) disagreement with these authors about the use of the term "field experiment." Field experiments have been the passion of my career. I have invested considerable effort in persuading other economists that experiments in real markets with real goods are an important technique for learning about the world. It wasn't easy, at first, to get people to take the work seriously. Preston McAfee, who recruited me recently to Yahoo! Research, confessed that when I was doing auction field experiments in graduate school, he and many others thought that I was "just playing games." He said that it took him a number of years to recognize the importance of field experiments, but he is now a great admirer of this research technique. I feel gratified that economists have now begun to see field experiments as an important part of the economist's toolkit.

I disagree with the usage, by these authors among others, of "field experiments" to refer to induced-value experiments that were performed using a portable laboratory in the "field" rather than a stationary laboratory at a university. What Harrison et al. call "field experiments," I prefer to call "lab experiments with subject pools other than university students," the same style of experiment that Dyer, Kagel, and Levin did with construction bidders back in 1989. I think this style of research is valuable, but I felt that the nomenclature diluted the term "field experiment," which I, and others, had been promoting as experiments with real transactions in naturally occurring markets.

My esteemed colleague John List, who largely shares my sentiments, coauthored with Harrison an influential paper categorizing types of field experiments (Harrison and List, 2004). I gather that they initially disagreed strongly over how to define a field experiment, but they finally settled (compromised?) on a taxonomy of three main types of field experiments. "Artifactual field experiments" refers to what I call portable-lab experiments, while "natural field experiments" refers to the sorts of studies I call field experiments. "Framed field experiments" are intermediate between the two, such as an experiment involving bidding on a commercial auction site for artificial goods with values induced by the experimenter.

Harrison and List have been quite influential with their taxonomy, as I now see many papers published with Harrison and List's "natural field experiment" nomenclature in their titles. However, I disagree with the nomenclature. First, I think that "artefactual field" and "natural field" are too cumbersome, when we could easily replace them with "field experiment" and "portable-lab experiment" (or "lab with Danish subjects aged 25–75," as in Harrison et al. (2002)). Second, I think that these two research strategies have fundamentally different intentions, and to use the same term to describe them only serves to confuse matters. A "natural field experiment" is designed to go into an existing market and make a manipulation that generates exogenous variation, in order to solve the problems of observational economists where Nature has not provided a good experiment. These also include various policy experiments (on educational policy, crime, job-training programs, etc.). An "artifactual field experiment" is designed to take protocols developed in the lab and experiment on different subject pools in order to explore robustness and document heterogeneity. Both share a commitment to using interventions to generate useful data, but they differ greatly in technique and intended purpose. Using "field experiment" to describe both activities increases confusion, and it requires people to use extra words to reduce the confusion. (Unfortunately, "natural field experiment" also runs the risk of confusion with "natural experiment," despite the important distinction that a field experiment involves deliberate treatment assignment by the researcher while a natural experiment cleverly takes advantage of a situation where Nature ran the experiment for us.)

Third, I disagree about the decision to lay out three categories in the taxonomy, instead of just making a contrast between two styles. What I love about the discussion of "framed field experiments" is the explicit observation that there are

many ways to be intermediate between a field experiment and a lab experiment: task, context, institutions, induced values, and even whether or not the subjects know they are being experimented on. All of these various features of the transactions are worth exploring for the different things they can teach us in the trade-off between realism and control. What I dislike is the attempt to make arbitrary distinctions to delineate the boundaries of the intermediate category. For example, I like to think of my dissertation research on auctions as a canonical example of a field experiment, but the Harrison–List taxonomy dictates calling that work a "framed field experiment"—it doesn't qualify for the full "natural field experiment" distinction because the MIT Human Subjects Committee required me to let all my online-auction bidders know that they were in an experiment. The variety of ways that experiments can get classified into the "framed field experiments" category makes them very different from each other, which makes that category less than useful. In my view, the concept of a continuum between laboratory and field experiments is really valuable, particularly as we note that the continuum has multiple dimensions to it. What's not helpful is trying to put some papers into a precise intermediate box, because that box contains more differences than similarities, in my opinion. This further muddies the water about what a field experiment is really about.

Despite the relatively wide acceptance of the Harrison–List taxonomy, I prefer to call my experiments either "field experiments" (most of my work) or "laboratory experiments with professional soccer players" (Levitt, List, and Reiley 2010). To me, the important part of the work with soccer players is not that we went out into the field and played games with these players in their team locker room, but rather that we recruited laboratory subjects with a particular kind of experience in playing games. The important characteristics of my field experiments on auctions, charitable fundraising, and advertising is that we have been able to use them to generate exogenous data to estimate causal effects of various real-world policy changes (auction format, solicitation strategy, advertising intensity). For that reason, I think that "natural field experiments" are much more different from typical laboratory experiments than are "artifactual field experiments." I would like to return to a simple distinction between "field experiments" and "lab experiments." In this framework, "field experiments" would include all "natural field experiments" and most "framed field experiments," while "laboratory experiments" would include most "artifactual laboratory experiments" and perhaps a few "framed field experiments." Laboratory experiments with unusual subject pools, or various other innovations, could refer to themselves as such. There will be a gray area in the middle that defies clear classification, but in those cases I don't mind whether we call it a field experiment or a lab experiment. My goal with this proposal is to use terms that are simple, clear, and evocative. I think that labels matter, and I prefer to see them be as useful as possible.

Despite my disagreements about taxonomy, I can certainly find agreement with Harrison et al. when they say "whether these are best characterized as being lab experiments or field experiments is not, to us, the real issue: the key thing is to see

this type of experiment along a continuum taking one from the lab to the field, to better understand behavior." I completely concur. I like to think that we can use a simpler taxonomy without obscuring that important point, which I first learned from Harrison and List (2004). I do think that labels matter and are useful.

Now for the substance of the paper. First, I really like the iMPL technique for eliciting risk preferences. I think that this iterative procedure is a major advance over the previous technique of using a fixed set of lottery choices to infer risk preferences. Assuming that the subjects behave relatively consistently across choices, this technique gives us the ability to zoom in on a more precise measurement of an individual's risk attitudes.

Second, I consider it an important observation that consistent estimates of time preference depend on measurements of individuals' risk preferences. If we assume risk neutrality when estimating time preferences (as is common practice), we find that people are much more impatient than if we allow them to be risk averse, and estimate their risk aversion using lottery choices. By the way, as far as I can tell, the dependency only goes in one direction: we should be able to estimate risk preferences just fine without having to estimate time preferences. (I found the paper somewhat unclear in its discussion of this issue.) The paper also makes the important point that correctly estimating standard errors on the discount rate requires joint estimation of risk and time preference parameters, so that any uncertainty about risk preference will correctly result in larger estimated uncertainty in the discount rate.

Next, I would like to step back from the trees of structural preference estimation and take a look at the forest. What are we really measuring when we measure risk aversion with an experiment? Are we measuring utility over income in the experiment? Utility over annual income? Utility over different wealth levels? Though not explicitly assumed in this work, researchers often assume implicitly that different types of risk preferences are the same. It may well be that what we care about for policy questions is the risk aversion over large changes in permanent wealth; but since we can't easily manipulate permanent wealth in the laboratory, we extrapolate from laboratory experiments on income and assume the parameters we estimate are relevant in the situations we care about. I want us to be more aware of these assumptions we're making.

I agree strongly with a comment made by Lise Vesterlund earlier in this conference: economic theory's comparative statics are much easier to verify with experiments than are point predictions. Therefore, I believe that Harrison et al. are likely pushing too hard on structural estimates (point predictions) of risk preferences, assuming that we can go a very long way toward answering important welfare questions using measures elicited from experiments. Before I am willing to believe in the welfare estimates, I need some reassurance that the required extrapolation is valid. One kind of evidence that could convince me would be a demonstration that we can use estimates from lottery-choice experiments to predict any sort of real-world choice behavior. Can we use this sort of risk-preference

estimate to predict life insurance purchases? Lottery ticket purchases? Deductibles chosen in automobile insurance policies? Investment choices in retirement accounts? I have not yet seen any evidence that we can use laboratory data to predict real-world transactions. If such evidence could be collected, I would get much more excited about the preference-elicitation research program.

To put things slightly differently, Harrison et al. are quite devoted to structural estimation of preference parameters. I am always wary of structural estimation, because such research tends, in my opinion, to rely too heavily on the particular parametric models chosen. The two-parameter expo-power model of risk preferences must be wrong at some level, because it's only a model. We just don't know how good or bad an approximation it is to the truth. Even though we go to great lengths to estimate accurate standard errors given the model, we don't ever inflate our standard errors to the extent that we are uncertain about the model itself (probably because we don't yet have a good mathematical procedure for incorporating specification uncertainty).[1] Thus, all structural parameter estimates are based on a large degree of faith in the model, and thus I prefer to maintain a high degree of skepticism about the estimates. I would be persuaded that we're really measuring the right thing if we could take structural point estimates and validate them by testing some kind of out-of-sample prediction. In this case, that could look like using structural estimates of risk and time preferences from laboratory-style elicitation experiments and using them to predict behavior in insurance choices, investment choices, education choices, and so on. Perhaps some future research program will employ clever field experiments (the sort that involve natural transactions) to test the structural estimates from laboratory-style experiments on the same individuals and see how well we can use them to extrapolate to other types of economic decision-making.

Final Observations: Theory Testing, Measurement, and Research Complementarities

What do field experiments contribute to theory testing? As noted above, laboratory experiments generally take a theory very seriously, impose all its primitives (independent private values, convex costs, etc.), and investigate whether behavior is consistent with predictions. I believe that laboratory experimentalists sometimes lose sight of the fact that theory is an abstraction and that the theory can miss a crucial detail that may change predictions. For example, when a charity raises money in a capital campaign for a threshold public good, will they really burn the money if they fail to reach their stated goal? Does auction theory accurately model bidder behavior on eBay, or does the existence of multiple competing auctions cause bidders to

bid differently? In a field experiment, we jointly test the primitives and the decision making process, to see how accurately a theory's predictions hold in the real world.

What do field experiments contribute to measurement and policy? A particularly brilliant idea in development economics, promoted by Michael Kremer and Esther Duflo, has been to randomize proposed policy treatments in order to figure out what actually works to promote health, education, income, and so on. It is hard for me to imagine an economic laboratory experiment being able to measure credibly the effects of medically de-worming schoolchildren on their educational attainment (Miguel and Kremer, 2003), or to what extent providing free mosquito netting can improve children's health outcomes (Cohen and Dupas, 2010).

Field experiments are clearly not a replacement for laboratory experiments on these topics. Rather, they open up new lines of inquiry on important economic questions that would be difficult to answer with a laboratory experiment. I can think of many other microeconomic examples of this sort. What is the optimal tax rate on cigarettes? How much does advertising affect consumer purchases? How do consumers react to prices with the last digit 9 versus the last digit 0? What is the elasticity of labor supply with respect to wages?

Often there will be productive interaction between laboratory and field experiments; one illustrative example comes from my own work. In Lucking-Reiley (1999), I used online field experiments to test revenue equivalence between auction formats. By contrast with the laboratory experiments of Cox et al. (1982), I found that I raised more money in a Dutch (declining-price) auction than in a first-price auction. There were a number of differences between my field experiments and the previous laboratory experiments, and it was not clear which caused the difference in revenues. Katok and Kwasnica (2008), in a subsequent laboratory experiment, managed to isolate one of these differences and show that it could explain the results. In my online auctions, I ran very slow Dutch auctions in order to fit the institutional features of the market: prices decreased just once per day, instead of the six-second interval previously used in the laboratory. Katok and Kwasnica showed that they could reverse the revenue rankings of the two auction formats in the laboratory by slowing the Dutch auction down. Thus, the field experiment led to a new discovery based on real institutional features, and the laboratory experiment was able to isolate an important cause of the field results.

I wish to close by stating how much I agree with a very important sentiment expressed in both of the papers I have been asked to discuss. Field experiments and lab experiments are extremely complementary to each other. Laboratory experiments allow us to observe and manipulate far more variables in service of testing theory. Field experiments, by getting us involved in real-world transactions, help us measure economic behavior in more realistic contexts.

I believe I speak for most field experimentalists when I acknowledge my debt to laboratory experimentalists, who have taught us to think of economics as an experimental science rather than a merely observational one. I believe that field experiments have the potential to engage the entire economics profession in learning

more from experiments in both the field and the laboratory. For some time, laboratory experiments have successfully engaged theorists in exploring more realism in decision-making processes. Now, by demonstrating their ability to generate exogenous variation in real-world situations, field experiments are getting the attention of empirical economists who have largely ignored experiments in the past, in their assumption that "real" economics was only an observational science. I feel strongly that improved dialogue between experimentalists and econometricians will lead to a much deeper understanding of economic behavior, and I urge other experimentalists to join me in promoting that dialogue.

NOTES

1. To their credit, the authors do relax their model with an analysis of a mixture model. But even this requires structural assumptions that, to my mind, are untested. My hope is that experiments will eventually nail down the structural regularities that we can rely on as tools in future estimation, but I don't think we're even close to this point yet.

REFERENCES

Bandiera, O., I. Barankay, and I. Rasul. 2005. Social Preferences and the Response to Incentives: Evidence from Personnel Data. *Quarterly Journal of Economics* **120**:917–962.

Cohen, J. and P. Dupas. 2010. Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment. *Quarterly Journal of Economics* **125**:1–45.

Cox, J. C., V. L. Smith, and J. M. Walker. 1982. Tests of a Heterogeneous Bidders Theory of First Price Auctions. *Economics Letters* **12**:207–212.

Dyer, D. and J. H. Kagel. 1996. Bidding in the Common Value Auctions: How the Commercial Construction Industry Corrects for the Winner's Curse. *Management Science* **42**:1463–1475.

Dyer, D., J. H. Kagel, and D. Levin. 1989. A Comparison of Naïve and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis. *Economic Journal* **99**:108–115.

Gneezy, U. and J. List. 2006. Putting behavioral Economics to Work: Field Evidence of Gift Exchange. *Econometrica* **74**:1365–1384.

Harrison, G. W. and J. A. List. 2004. Field Experiments. *Journal of Economic Literature* **42**:1009–1055.

Harrison, G. W., M. Lau, and E. E. Rutstrom. 2015. Theory, Experimental Design and Econometrics Are Complementary (And So Are Lab and Field Experiments). In *Handbook of Experimental Economic Methodology,* eds. G. Fréchette and A. Schotter, pp. 296–338. New York: Oxford University Press.

Harrison, G. W., M. Lau, and M. Williams. 2002. Estimating Individual Discount Rates in Denmark: A Field Experiment. *American Economic Review* **92**:1606–1617.

Kagel, J. H. 2015. Laboratory Experiments. In *Handbook of Experimental Economic Methodology,* eds. G. Fréchette and A. Schotter, pp. 339–359. New York: Oxford University Press.

Katok, E. and A. M. Kwasnica. 2008. Time is Money: The Effect of Clock Speed on Seller's Revenue in Dutch Auctions. *Experimental Economics* **11**:344–357.

Levitt, S. D, J. A. List, and D. Reiley. 2010. What Happens in the Field Stays in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments. *Econometrica* **78**:1413–1434.

Lucking-Reiley, D. 1999. Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet. *American Economic Review* **89**:1063–1080.

Miguel, E. and M. Kremer. 2003. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* **72**:159–217.

Palacios-Huerta, I. and O. Volij. 2008. Experientia Docet: Professionals Play Minimax in Laboratory Experiments. *Econometrica* **76**:75–115.

Wason, P. C. 1966. Reasoning. In *New Horizons in Psychology*, ed. B. M. Foss. Harmondsworth: Penguin.

CHAPTER 20

ON THE GENERALIZABILITY OF EXPERIMENTAL RESULTS IN ECONOMICS

OMAR AL-UBAYDLI
AND JOHN A. LIST

Introduction

> The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past. (Knight, 1921, p. 313)

More than 15 years ago, one of the coauthors (List) sat in the audience of a professional presentation that was detailing whether and to what extent students collude in the lab and what this meant to policy makers interested in combating collusion. He openly wondered how such behavior would manifest itself with live traders in an extra-lab market, asking innocently whether policy makers should be concerned that this environment was much different from the one in which they typically operate. His concerns were swept aside as naive.

Later in that same year, List attended a conference where experimental economists debated the merits of an experimental study that measured the

magnitude of social preferences of students. He asked if such preferences would thrive in naturally occurring settings and how they would affect equilibrium prices and quantities. In not so many words, he was told to go and sit in the corner again. After the session, another junior experimentalist approached a now distraught List: "Those are great questions, but off limits." List queried why, to which he received a response "That's the way it is."[1]

Except for the names and a few other changes, List was articulating words in the spirit of what Knight had eloquently quipped nearly 100 years ago: The intriguing possibility of using laboratory experiments as a solution to real-world problems depended on the lab being like the field in terms of delivering similar behavioral relationships. A wet behind the ears List was fascinated by this query, but was learning that others did not share his passion, or even his opinion that it was a worthwhile point to discuss.

We are happy to find that the good ol' days are behind us. Today it is not uncommon for the very best minds in economics to discuss and debate the merits of the experimental method and the generalizability of experimental results (e.g., Falk and Heckman, 2009; Camerer, Chapter 14, this volume; Fréchette, Chapter 17, this volume; Kessler and Vesterlund, Chapter 18, this volume). We find this fruitful for many reasons, and we continue to scratch our heads when some critics continue to contend that we have "ruined the field of experimental economics' by scribing the original Levitt and List (2007b; henceforth LL) article. This is a very short run view; indeed, our field of experimental economics can be sustainable only if our audience includes those outside our direct area of study. Otherwise, we run the real risk of becoming obscure. Understanding the applicability of our empirical results and having an open discussion can move us closer to the acceptance of our tools by all economists, and it can move us toward an approach that can help us more fully understand the economic science.

More broadly, this volume represents a sign of change—we have entered a climate of scientific exploration that permits a serious investigation of what we believe to be the most important questions facing behavioral and experimental economists: (1) Which insights from the lab generalize to the extra lab world? (2) How do market interactions or market experience affect behaviors? (3) Do individual behaviors aggregate to importantly affect market equilibria, and how does equilibration affect the individual behaviors?

The object of this forum is to discuss the recent study due to LL. For the most part, the critics writing in this volume understood LL's contributions and hypotheses, and they wrote a balanced and thoughtful evaluation of that work. As a point of reference, one of LL's contributions was to present a theoretical framework and gather empirical evidence that questioned the level, or point, estimates delivered by laboratory experiments in economics. As a point of discussion, they focused on the work within the area of the measurement of social preferences. LL's overarching points included arguments that the laboratory is especially well equipped to deliver qualitative treatment effects, or comparative static insights, but not well suited to

deliver deep structural parameters, or precise point estimates. This is because such estimates critically depend on the properties of the situation, as they detailed with examples from economics and psychology experiments.

In the end, LL argue that lab and field experiments are complements with each serving an important role in the discovery process (consistent with what List has argued in all of his work, as pointed out by some of the commentators in this volume). We suspect that the commentators are in full agreement with this broader point.

In this study we begin by providing an overview of experimental methods in economics, focusing on the behavioral parameters that each estimates. We then turn to formalizing generalizability. In principle, generalizability requires no less of a leap of faith in conventional (non-experimental) empirical research than in experimental research. The issue is obfuscated in non-experimental research by the more pressing problem of identification: how to correctly estimate treatment effects in the absence of randomization.

In our model, we generalize the 'all causes' approach to a more continuous form where researchers have priors about causal effects and update them based on data. This formality is necessary for a precise articulation of a theory of the advantages offered by field experiments. We then place the theory into a convenient "updating" model that shows the power of replication: We argue that just a few replications yields a tremendous increase in the probability that the received result is in fact true. Our penultimate section addresses some of the various criticisms leveled by the discussants. We conclude with some thoughts on where we hope this line of research goes in the coming years.

Preamble: Empirical Methods

The empirical gold standard in the social sciences is to estimate a causal effect of some action. For example, measuring the effect of a new government program or answering how a new innovation changes the profit margin of a firm are queries for the scientist interested in causal relationships. The difficulty that arises in establishing causality is that either the action is taken or it is not—we never directly observe what would have happened in an alternative state in which a different action is taken. This, combined with the fact that in the real world there are simultaneously many moving parts, has led scholars to conclude that experimentation has little hope within economics.

Such thoughts reflect a lack of understanding of how the experimental method identifies and measures treatment effects. In fact, complications that are difficult to understand or control represent key reasons to *conduct* experiments, not a point of skepticism. This is because randomization acts as an instrumental variable, balancing unobservables across control and treatment groups.

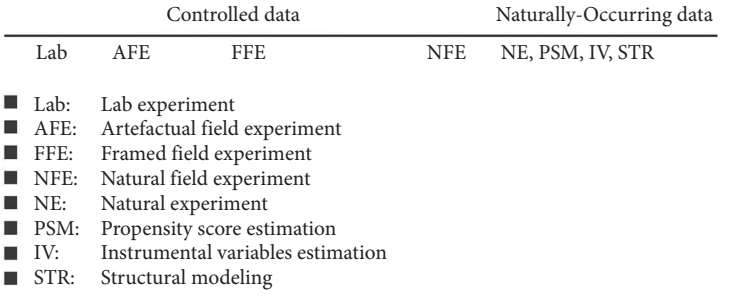


Figure 20.1. A field experiment bridge.

To show this point, we find it instructive to consider empirical methods more broadly. The easternmost portion of Figure 20.1, which we have often used elsewhere, highlights some of the more popular approaches that economists use to analyze naturally occurring data.

For example, identification in natural experiments results from a difference-in-difference (DD) regression model where the major identifying assumption is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with treatment status and that selection into treatment is independent of the temporary individual specific effect. For example, let's say that the researcher is interested in estimating the impact on labor supplied from an increase in minimum wage, as Card and Krueger (1994) famously do by comparing labor supplied at fast food restaurants in New Jersey—which raised their minimum wage—and neighboring Pennsylvania—which did not change their minimum wage. There's no ex ante reason to expect New Jersey and Pennsylvania to start with the same labor supplied, but the motivation behind using DD is that you would expect the difference in labor supplied from year to year in both states to be pretty similar, all else being equal.

Card and Krueger leverage the policy change in New Jersey to compare the difference of those differences in order to understand the impact of minimum wage laws on the quantity of labor supplied. Implicit in their analysis, though, is that other than the change in minimum wage laws in New Jersey, nothing has impacted the difference in the quantity of labor supplied between the time periods in Pennsylvania that is correlated with treatment. Furthermore, they must assume that treatment was randomly applied to New Jersey and not Pennsylvania, otherwise we don't know whether New Jersey just has some unique trait that is correlated with treatment status that would impact the quantity of labor supplied.

Useful alternatives to this approach include the method of propensity score matching (PSM) developed in Rosenbaum and Rubin (1983). A major assumption under this approach is called the "conditional independence assumption," and it intuitively means that selection into treatment occurs only on observables. This means, for example, that the econometrician knows all the variables that influence whether a person selects into an employment program. In most cases,

this assumption is unrealistic. Other popular methods of measurement include the use of instrumental variables and structural modeling. Assumptions of these approaches are well documented and are not discussed further here (see, e.g., Rosenzweig and Wolpin (2000) and Blundell and Costa Dias (2002)).

We think that it is fair to say that these approaches of modeling naturally occurring data are quite useful; but because the world is complicated, they are sometimes subject to incredulous assumptions. We are not the first to make this point, as there are entire literatures discussing the limitations of the various empirical models. In essence, many people argue that because the economic world is extremely complicated, one must take great care when making causal inference from naturally occurring data.

In the westernmost portion of Figure 20.1 is the laboratory experiment, which typically makes use of randomization to identify a treatment effect of interest among student subjects. Making generalizations outside of this domain might prove difficult in some cases, but to obtain the effect of treatment in this particular domain the only assumption necessary is appropriate randomization.

Field experiments represent a movement to take the data generation process beyond the walls of the laboratory. Two decades ago, the primary data generators were lab experimentalists. The past 15 years has witnessed an explosion of creative ways to generate data in the field. Harrison and List (2004) propose six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate. Using these factors, they discuss a classification scheme that helps to organize one's thoughts about the factors that might be important when moving from the lab to the field.

According to this classification scheme, the most minor departure from the typical laboratory experiment is the "artifactual" field experiment (AFE), which mimics a lab experiment except that it uses "nonstandard" subjects. Such subjects are nonstandard in the sense that they are not students, but instead participants drawn from the market of interest. This type of experiment represents a useful type of exploration beyond traditional laboratory studies, as explored in Fréchette (Chapter 17, this volume). As Fréchette discusses, AFEs have been fruitfully used in financial applications, public economics, environmental economics, and industrial organization, as well as to test predictions of game theory.

Moving closer to how naturally occurring data are generated, Harrison and List (2004) denote a framed field experiment (FFE) as the same as an AFE but with field context in the commodity, task, stakes, or information set that the subjects can use. This type of experiment is important in the sense that a myriad of factors might influence behavior; and by progressing slowly toward the environment of ultimate interest, one can learn about whether, and to what extent, such factors influence behavior one by one.

FFEs represent a very active type of field experiment in the past decade. Social experiments and recent experiments conducted in development economics are a type of FFE: Subjects are aware that they are taking part in an experiment, and in many cases they understand that their experience is for research purposes. Peter Bohm was an early experimenter to depart from traditional lab methods by using FFEs (Bohm, 1972). While his work touched off an interesting stream of research within environmental and resource economics, for a reason that we cannot quite put our finger on, the broader economics literature did not quickly follow Bohm's lead to pursue research outside of the lab. This has only happened in the past decade or so.

Finally, a natural field experiment (NFE) is the same as a FFE in that it occurs in the environment where the subjects naturally undertake these tasks, but where the subjects *do not know* that they are participants in an experiment.[2] Such an exercise is important in that it represents an approach that combines the most attractive elements of the experimental method and naturally occurring data: randomization and realism. In addition, it importantly tackles a selection problem that is not often discussed concerning the other types of experiments, as discussed below.

NFEs have recently been used to answer a wide range of questions in economics, including topics as varied as measuring preferences (List, 2003) and how one can manage an on-line shopping experience (Hossain and Morgan, 2006). The economics of charity has witnessed a plethora of NFEs, as recently discussed in List (2011a). Of course, the taxonomy in Figure 20.1 leaves gaps, and certain studies may not fall neatly into such a classification scheme, but such an organization highlights what is necessary in terms of scientific discovery to link controlled experimentation to naturally occurring data.

As we will argue below, an NFE represents the cleanest possible manner in which to estimate the treatment effect of interest. In this light, economists can certainly go beyond activities of astronomers and meteorologists and approach the testing of laws akin to chemists and biologists. Importantly, however, background variables can matter greatly when one attempts to generalize empirical results. With an understanding of the exact behavioral parameters identified by the various experimental approaches, we will be in a position to discuss generalizability, the focus of this volume. We first turn to the estimated parameters from experiments.

What Parameters Do Experiments Estimate?

Without loss of generality, define y_1 as the outcome with treatment, with y_0 being the outcome without treatment. The treatment effect for person i can then be measured as $\tau_i = y_{i1} - y_{i0}$. The major problem, however, is one of a missing counterfactual—person i is not observed in both states of the world. We assume that $p = 1$ indicates participation in the experiment, whereas $p = 0$ indicates nonparticipation. That is, people who agree to enroll in the experiment have $p = 1$,

others have $p = 0$. In this way, if one is interested in the mean differences in outcomes, then the treatment effect of interest is given by

$$t = E(\tau | p = 1) = E(y_1 - y_0 | p = 1).$$

Yet, in our experience in the field, what is typically reported by government programs such as *Head Start*, firms (non-profits and for profits) and laypeople who discuss results from experiments, is a treatment effect as follows:

$$t' = E(y_1 | p = 1) - E(y_0 | p = 0).$$

Such a reported effect represents a potentially misleading measurement because it is comparing the mean outcome for two potentially quite different populations. To see the difference between t and t', simply add and subtract $E(y_0 | p = 1)$ from t', yielding

$$t' = \underbrace{E(y_1 - y_0 | p = 1)}_{t} + \underbrace{E(y_0 | p = 1) - E(y_0 | p = 0)}_{\delta}$$

where δ is the traditional selection bias term. This bias is a result of the nontreated differing from one another in the *nontreated state*.

This equation is illustrative because it shows clearly how selection bias, as is typically discussed in the literature, relates to outcomes in the nontreated state. For example, if parents who care more deeply about their children's educational outcomes are those who are more likely to sign up for services from *Head Start*, then their children might have better outcomes in the nontreatment state than children of parents who care less deeply about their children's educational outcomes. In this case, such selection bias causes the second term to be greater than zero because $E(y_0 | p = 1) > E(y_0 | p = 0)$, leading the *Head Start* program to report (a) a treatment effect that is too optimistic or (b) a treatment effect estimate that is biased upwards. In such instances, we would systematically believe that the benefits of *Head Start* are considerably higher than their true benefits. In our travels, we have found that this problem—one of not constructing the proper control group—is ubiquitous.

To avoid this sort of selection bias, what is necessary is for randomization and identification of the treatment effect to occur just over the $p = 1$ group, yielding a treatment effect estimate of the mean outcome differences between treated and nontreated from the $p = 1$ group. Letting $D = 1$ (0) denote those randomized into treatment (nontreatment):

$$t = E(y_1 | D = 1 \text{ and } p = 1) - E(y_0 | D = 0 \text{ and } p = 1).$$

At this point, it is instructive to pause and ask how to interpret the meaning of this treatment effect. First, this is the treatment effect that laboratory experiments,

as well as AFEs and FFEs, report (but not the treatment effect reported from NFEs). Given that randomization was done appropriately, this is a valid treatment effect estimate for the $p = 1$ population. For this effect to generalize to the $p = 0$ population, however, further assumptions must be made.

For example, the effect of treatment cannot differ across the $p = 1$ and $p = 0$ groups. If, for instance, a person has a unique trait that is correlated with treatment status and correlated with the outcome variable, such generalization is frustrated. In our *Head Start* example, it might be the case that parents who believe *Head Start* will have a positive effect on their child are more likely to enroll. In that case, it would not be appropriate to generalize the effect from the $p = 1$ group to the $p = 0$ group if such beliefs were actually true.

This effect—call it treatment-specific selection bias—is quite distinct from the traditional selection bias discussed in the literature and shown above. Whereas the standard selection bias relates to outcomes of the $p = 1$ and $p = 0$ groups in the nontreated state, this sort of bias in the measured treatment effect related to outcomes of the $p = 1$ and $p = 0$ groups in the *treated* state.

So how do NFEs differ in their identification approach? Since subjects are not aware that they are taking part in an experiment, NFEs naturally resolve any bias issues. In this case, there is no $p = 1$ or $p = 0$ group: Subjects are randomly placed into treatment or control groups without even knowing it. This fact excludes the typical selection effect discussed in the literature and precludes treatment-specific selection bias. Indeed, it also rids us of other biases, such as randomization bias and any behavioral effects of people knowing that they are taking part in an experiment.

The very nature of how the parameter is estimated reveals the mistake that many people make when claiming that the laboratory environment offers more "control" than a field experiment. There are unobservables in each environment, and to conclude *ex ante* that certain unobservables (field) are more detrimental than others (lab) is missing the point. This is because randomization balances the unobservables—whether a myriad or one. Thus, even if one wished to argue that background complexities are more severe in one environment than in the other, there really is little meaning—one unobservable can do as much harm as multiple unobservables. Indeed, all it takes is for one unobservable to be correlated with the outcome for an approach to have a problem of inference. The beauty behind randomization is that it handles the unobservability problem, permitting a crisp estimate of the causal effect of interest.

Formalizing Generalizability

When we first began to explore generalizability, we found a dearth of theory and smattering of empirical evidence.[3] Even though we presented a theoretical framework in LL, our attention there was focused on the empirical evidence.

Accordingly, here we focus on the theory and leave it to the interested reader to scrutinize the extant literature and make an informed opinion about what it says. Our own opinion is that it is too early to tell decisively where the empirical debate will end, but the evidence is mounting in favor of the hypotheses in LL. But, as usual, caveat lector—we leave it to the reader to decide.

In the all-causes model (Heckman, 2000), the researcher starts with a causal effect about which she has no prior. The purpose of an empirical investigation is to generate an estimate. In this section, we will generalize the all-causes model to a more continuous form where researchers have priors about causal effects and update them based on data. This formality is necessary for a precise articulation of a theory of the advantages offered by field experiments; it is also consonant with our empirical complement presented below.

Setup

Let Y be a random variable, denoted the **dependent variable**, whose realizations are in $S_Y \subseteq \mathbb{R}$; let X be a random variable, denoted the **explanatory variable of interest**, whose realizations are in $S_X \subseteq \mathbb{R}$; and let Z be a random vector, denoted the **additional explanatory variables**, whose realizations are in $S_Z \subseteq \mathbb{R}^k$. Furthermore, Z contains all the explanatory variables (apart from X) that have an impact on Y. To focus our model on the generalizability problem (rather than the sampling/inference problem), we assume that Z is observable. This model can be easily expanded to allow for unobservable variables.

In the all causes model, (X, Y, Z) are related according to the function $f : S_X \times S_Z \rightarrow S_Y$. Each $\left(x, x', z\right) \in S_X \times S_X \times S_Z$ is denoted a **causal triple**. The **causal effect** of changing X from x to x' on Y given $Z = z$ is described by the function $g : S_X \times S_X \times S_Z \rightarrow \mathbb{R}$, where

$$g\left(x, x', z\right) = f\left(x', z\right) - f(x, z).$$

Let $T \subseteq S_X \times S_X \times S_Z$ be the **target space**. It describes the causal triples in which an empirical researcher is interested. Typically, she wants to know the exact value of the causal effect, $g\left(x, x', z\right)$, of each element of T. Often, particularly in experimental research, a researcher is interested merely in knowing if the causal effect lies in a certain range. Let $h : S_X \times S_X \times S_Z \rightarrow \mathbb{R}$ be a function that captures the aspect of a causal effect in which the researcher is interested. The most common, especially when testing theory (rather than selecting policy), is $\overline{h}$:

$$\overline{h}\left(x, x', z\right) = \begin{cases} -1 & \text{if } g\left(x, x', z\right) < 0, \\ 0 & \text{if } g\left(x, x', z\right) = 0, \\ 1 & \text{if } g\left(x, x', z\right) > 0. \end{cases}$$

Before embarking upon a new empirical investigation, a researcher has a **prior** $\mathcal{F}^0_{x,x',z} : \mathbb{R} \rightarrow [0, 1]$ about the value of $h\left(x, x', z\right)$ for each $\left(x, x', z\right) \in T$. The prior is

a cumulative density function based on existing theoretical and empirical studies, as well as researcher introspection.

An empirical investigation is a **dataset** $D \subseteq S_X \times S_X \times S_Z$. Note that D and T may be disjoint, and both may be singletons. Indeed, D is often a singleton in laboratory experiments. The researcher will typically sample Y repeatedly at $(X, Z) = (x, z)$ and $(X, Z) = (x', z)$ and use this to obtain an estimate of $g\left(x, x', z\right)$. Let the **results** $R \subseteq D \times \mathbb{R}$ be the set of causal effects obtainable from the dataset D *making no parametric assumptions* (i.e., no extrapolation or interpolation):

$$R = \left\{\left(x, x', z, g\left(x, x', z\right)\right) : \left(x, x', z\right) \in D\right\}.$$

As mentioned above, we set aside the sizeable problem of obtaining a consistent estimate of $g\left(x, x', z\right)$. In fact this is the primary problem faced by most non-experimental, empirical research due to, for example, small samples and endogeneity problems. To some extent, generalizability is a secondary issue in empirical research that uses naturally occurring data simply because it is overshadowed by the more pressing issue of identification.

This essay will ignore this part of the identification problem to focus attention upon the generalizability problem. Questions about how sample size and variance affect the estimation procedure are set aside because they do not interact with the main principles, though this framework can be easily expanded to incorporate such issues. Consequently, we do not draw a distinction between a causal effect $g\left(x, x', z\right)$ and a direct empirical estimate of $g\left(x, x', z\right)$.

After seeing the results, R, the researcher updates her prior $\mathcal{F}^0_{x,x',z}$ for each $\left(x, x', z\right) \in T$, forming a **posterior** $\mathcal{F}^1_{x,x',z}$. *The updating process is not necessarily Bayesian.* The generalizability debate, which we discuss in the next section, is concerned with the formation of the posterior, especially for elements of $T \backslash D$. We henceforth assume that the prior is never completely concentrated at the truth, implying that any valid estimate of $g\left(x, x', z\right)$ will always lead to the researcher updating her prior.

The posterior is the conclusion of the empirical investigation. This framework is designed to include studies that estimate causal effects for policy use, for testing a theory, or for comparing multiple theories.

To put the framework into motion with an economic example, we consider a Laffer-motivated researcher who wants to know if increasing sales tax (X) from 10% to 15% increases tax revenue (Y) when the mean income in a city (Z) is \$30k. For expositional simplicity, we assume that the only element of Z is mean income level. The researcher can only generate data in four cities: Two cities have a mean income of \$20k and two cities have a mean income of \$35k. All four cities currently have a sales tax of 10%. She randomly assigns treatment (increasing sales tax to 15%) to one city in each income pair and control (leaving the sales tax at 10%) to the other city in each pair. She then collects data on tax revenue (one observation in each cell is sufficient because we are not tackling the sample-size component of the identification problem).

The researcher's prior is a 0.5 chance of a positive causal effect at a mean income of $30k. She finds a positive causal effect at both mean income levels and revises her prior at a mean income of $30k to a 0.6 chance of a positive causal effect. In terms of our notation, we have

$$T = \{(10\%, 15\%, \$30,000)\},$$

$$h\left(x, x', z\right) = \begin{cases} 1 & \text{if } g\left(x, x', z\right) > 0, \\ 0 & \text{if } g\left(x, x', z\right) \leq 0, \end{cases}$$

$$D = \{(10\%, 15\%, \$20000), (10\%, 15\%, \$35000)\},$$

$$R = \{(10\%, 15\%, \$20000, 1), (10\%, 15\%, \$35000, 1)\},$$

$$\mathcal{F}^0_{10\%,15\%,\$30,000}(0) = 0.5, \mathcal{F}^1_{10\%,15\%,\$30000}(0) = 0.4.$$

Different Types of Generalizability

Given a set of priors $\mathcal{F}^0 = \{\mathcal{F}^0_{x,x',z} : \left(x, x', z\right) \in S_X \times S_X \times S_Z\}$ and results R, the **generalizability set** $\Delta(R) \subseteq \{S_X \times S_X \times S_Z\} \backslash D$ is the set of causal triples outside the dataset where the posterior $\mathcal{F}^1_{x,x',z}$ is updated as a consequence of learning the results:

$$\Delta(R) = \{\left(x, x', z\right) \in \{S_X \times S_X \times S_Z\} \backslash D : \mathcal{F}^1_{x,x',z}(\theta) \neq \mathcal{F}^0_{x,x',z}(\theta) \text{ for some } \theta \in R\}.$$

Results are **generalizable** when the generalizability set is non-empty: $\Delta(R) \neq \emptyset$. A researcher is said to **generalize** when the generalizability set intersects with the target space: $\Delta(R) \cap T \neq \emptyset$. The researcher in the above Laffer example is generalizing. Note that generalizability is focused on $h\left(x, x', z\right)$ rather than $g\left(x, x', z\right)$ since the prior is focused on $h\left(x, x', z\right)$.

As mentioned above, in principle, generalizability requires no less of a leap of faith in conventional (non-experimental) empirical research than in experimental research. The issue is obfuscated in non-experimental research by the more pressing problem of identification: how to correctly estimate $g\left(x, x', z\right)$ in the first place due to, for example, the absence of randomization. This problem does not plague experimental work. Indeed, the beauty of experimentation is that through randomization the problem of identification is solved.

Given prior beliefs $\mathcal{F}^0$, a set of results R has **zero generalizability** if its generalizability set is empty: $\Delta(R) = \emptyset$. Zero generalizability is the most conservative empirical stance and equates to a paralyzing fear of interpolation, extrapolation, or the assumption of additive separability.

Given prior beliefs $\mathcal{F}^0$, a set of results R has **local generalizability** if its generalizability set contains points within an arbitrarily small neighborhood of points in D:

$$\left(x, x', z\right) \in \Delta(R) \Rightarrow \left(x, x', z\right) \in B_\varepsilon\left(x, x', z\right) \qquad \text{for some } \varepsilon > 0, \left(x, x', z\right) \in D.$$

The simplest way to obtain local generalizability is to assume that $h\left(x, x', z\right)$ is continuous (or only has a small number of discontinuities), since continuity implies local linearity and therefore permits local extrapolation.[4] In the Laffer example above, assuming that the causal effect is continuous in the mean income level in the city, the researcher can extrapolate her findings to estimate the causal effect for a city with a mean income level of $35100. In principle, nonlocal changes in $\left(x, x', z\right)$ can have a large effect on h, limiting our ability to extrapolate. However, *as long as we do not change* $\left(x, x', z\right)$ *by much and* $h\left(x, x', z\right)$ *is continuous, then* h *will not change by much* and so our dataset D will still be informative about causal effects outside this set.

Since continuity is sufficient for local generalizability, it follows that discontinuity is necessary for zero generalizability. If, as is often likely to be the case, the researcher is unsure of the continuity within $h\left(x, x', z\right)$, then the more conservative she is, the more she will be inclined to expect zero generalizability.[5]

Given prior beliefs $\mathcal{F}^0$, a set of results R has **global generalizability** if its generalizability set contains points outside an arbitrarily small neighborhood of points in D:

$$\exists \left(x, x', z\right) \in \Delta(R) : \left(x, x', z\right) \notin B_\varepsilon(x, x', z) \quad \text{for some } \varepsilon > 0, \text{ for all } (x, x', z) \in D.$$

In the Laffer example above, the researcher is assuming global generalizability. At its heart, *global generalizability is about assuming that a large change in* $\left(x, x', z\right)$ *does not have a large effect on* h.

A succinct summary of the section entitled "Formalizing Generalizability" thus far is as follows.

1. In a nonparametric world, results can fail to generalize, generalize locally, or generalize globally.
2. A sufficient condition for local generalizability is continuity of $h\left(x, x', z\right)$.
3. A sufficiently conservative researcher is unlikely to believe that her results generalize globally because this requires a much stronger assumption than continuity.

We are now in a position to formalize the advantages offered by field experiments.

A Theory of the Advantage Offered by Field Experiments

A (function of a) causal effect $h\left(x, x', z\right)$ is **investigation neutral** if it is unaffected by the fact that it is being induced by a scientific investigator *ceteris paribus*. Thus, for example, suppose that we are studying the causal effect of the slope of a demand curve on the percentage of surplus realized in a market. If this effect is investigation neutral, then the fact that the market was set up as the result of a scientific investigation versus simply being observed in the naturally occurring domain,

ceteris paribus, does not change the causal effect. **We assume that causal effects are investigation neutral.**

We define a **natural setting** as a triple (x, x', z) that can plausibly exist in the absence of academic, scientific investigation. For example, if a scientist is studying the effect of a piece-rate versus a fixed-wage compensation scheme on the productivity of a worker soliciting funds in a phoneathon for a charity, then this is a natural setting since it is common for workers to get hired to do such tasks using a piece-rate or a fixed-wage scheme. In contrast, if a scientist is interested in studying the magnitude of social preferences and brings a group of students into the lab to play a dictator game, then this is not a natural setting since students virtually never find themselves involved in such a scenario under the specific features of that environment and task.

Our principal assumption is that as economists, we are more interested in learning about understanding behavior in natural settings than in non-natural settings. This does not eliminate the value of learning about causal effects in non-natural settings; after all, the benefits of centuries of artificial studies in physics, chemistry, and engineering are self-evident. However, it requires that insights gained in non-natural settings generalize to natural settings for them to be of great value. This is because as economists we are interested in reality, in contrast to, say, poetry. We are concerned with understanding the real world and in modifying it to better the allocation of scarce resources or to prescribe better solutions to collective choice problems.

Through this lens, because of their very nature, laboratory experiments represent an environment that could only ever come about as the result of a scientific investigation. **Thus, laboratory investigations are not completed in natural settings. Moreover, many laboratory experiments might not even be in the neighborhood of a natural setting**. This is because several variables have to change by large amounts in order for a laboratory setting to transform into a natural setting—for example, the nature and extent of scrutiny, the context of the choice decision and situation, the experience of participants, and several other factors discussed in LL. We elaborate on one such factor—the participation decision—below.

Falk and Heckman (2009) and others (including Camerer, Chapter 14 in this volume) have questioned whether the nonlocal changes in (x, x', z) that arise when generalizing from a laboratory setting to a field setting have a large effect on $h(x, x', z)$. Interestingly, when making their arguments, they ignore one of the most important: Typical laboratory experiments impose artificial restrictions on choice sets and time horizons.

Regardless of the factors that they discuss and fail to discuss, to the best of our knowledge, nobody has questioned the proposition that the changes in (x, x', z) are nonlocal.[6] In fact, the artificial restrictions on choice sets and time horizons are a particularly dramatic illustration of the nonlocal differences between laboratory and field settings. Another critical, nonlocal difference between laboratory and

natural field settings is the participation decision, shown above in the traditional treatment effects model and discussed below within our framework.

With this background in hand, we proceed to three propositions, which are meant to capture the range of thoughts across the economics profession today. We do not believe that one can categorize all laboratory experiments under any one of these propositions, but rather believe that there are a range of laboratory experiments, some of which fall under each of the three propositions.

Proposition 20.1.

Under a liberal stance (global generalizability), neither field nor laboratory experiments are demonstrably superior to the other.

This view is the most optimistic for generalizing results from the lab to the field. It has as its roots the fact that the generalizability sets are both non-empty and, in general, neither will contain the other. In this way, empirical results are globally generalizable.

As an example, consider the work on market equilibration. Conventional economic theory relies on two assumptions: utility-maximizing behavior and the institution of Walrasian tâtonnement. Explorations to relax institutional constraints have taken a variety of paths, with traditional economic tools having limited empirical success partly due to the multiple simultaneously moving parts in the marketplace. Vernon Smith (1962) advanced the exploration significantly when he tested neoclassical theory by executing double-oral auctions. His results were staggering: Quantity and price levels were very near competitive levels after a few market periods. It is fair to say that this general result remains one of the most robust findings in experimental economics today.

List (2004) represents a field experiment that moves the analysis from the laboratory environment to the natural setting where the actors actually undertake decisions. The study therefore represents an empirical test in an actual marketplace where agents engage in face-to-face continuous bilateral bargaining in a multilateral market context.[7] Much like Smith's (1962) setup, the market mechanics in List's bilateral bargaining markets are not Walrasian.

Unlike Smith (1962), however, in these market subjects set prices as they please, with no guidance from a centralized auctioneer. Thus, List's design shifts the task of adaptation from the auctioneer to the agents, permitting trades to occur in a decentralized manner, similar to how trades are consummated in actual free unobstructed markets. In doing so, the market structure reformulates the problem of stability of equilibria as a question about the behavior of actual people as a psychological question—as opposed to a question about an abstract and impersonal market.

A key result of List's study is the strong tendency for exchange prices to approach the neoclassical competitive model predictions, especially in symmetric

markets. This example highlights exactly what the original LL model predicts: A wide class of laboratory results should be directly applicable to the field. In particular, we would more likely find an experiment falling under Proposition 20.1 when (a) the experimenter does not place the subject on an artificial margin, (b) moral concerns are absent, (c) the computational demands on participants are small, (d) nonrandom selection of participants is not an important factor, (e) experience is unimportant or quickly learned, and (f) the experimenter has created a lab context that mirrors the important aspects of the real-world problem. At that point, we would expect results from the lab to be a closer guide to natural settings.

Our next proposition strengthens this liberal view.

Proposition 20.2.

Under a conservative stance (local generalizability; or if the researcher is confident that $h\left(x, x', z\right)$ is continuous), field experiments are more useful than laboratory experiments.

This view follows from the idea that results are generalizable locally. Thus, whether empirical data are generated in the lab or in the field, it can be generalized to the immediately adjacent settings. And, since field experiments provide information from a natural setting and laboratory experiments from a non-natural setting, field experiment are more useful. This is because the neighborhood of a natural setting is still a natural setting, while the neighborhood of a non-natural setting is non-natural.

As an example, consider the recent work in the economics of charity. Without a doubt, the sector represents one of the most vibrant in modern economies. In the United States alone, charitable gifts of money have exceeded 2% GDP in the past decade. Growth has also been spectacular: From 1968 to 2008, individual gifts grew nearly 18-fold, doubling the growth rate in the S&P 500. Recently, a set of lab and field experiments have lent insights into the "demand side" of charitable fundraising.

For instance, consider the recent laboratory experiments of Rondeau and List (2008). They explored whether leadership gifts—whether used as a challenge gift (simply an announcement) or as a match gift (i.e., send in $100 and we will double your contribution)—affect giving rates. From the lab evidence, they found little support for the view that leadership gifts increase the amount of funds raised.

Alternatively, in that same paper, they used leadership gifts to raise money for the Sierra Club of Canada via a field experiment. Their natural field experiment was conducted within the spirit of one of the typical fundraising drives of the Sierra Club organization. A total of 3000 Sierra Club supporters were randomly divided into four treatments, varying the magnitude and type of leadership gift. They find that challenge gifts work quite well in the field. This means that it is important for fundraisers to seek out big donors privately before they go public with their cause, and to use challenge gifts when doing so.

One is now in a position to ask: If I am a fundraiser, which set of results should guide my decision making, those from the lab or those from the field?

Viewed through the lens of Proposition 20.2, practitioners in the field who are interested in raising money for their cause would be well served to pay close attention to the field experimental results because such insights are locally generalizable. On the other hand, the lab results that suggest that the upfront monies raised will not help much *are less likely* to generalize outside of the lab confines.

This result highlights that economists are often only concerned with obtaining the sign of a causal effect $g\left(x, x', z\right)$, as summarized by the function $\overline{h}\left(x, x', z\right)$ above. In this case, if the researcher is confident that $g\left(x, x', z\right)$ is monotonic in z_i over some range $[z_{i0}, z_{i1}]$, then $\overline{h}\left(x, x', z\right)$ will be continuous almost everywhere. This is sufficient for local generalizability.

Finally, an even further tightening of the restriction set leads to our third proposition.

Proposition 20.3.

Under the most conservative stance (zero generalizability), field experiments are more useful than laboratory experiments because they are performed in one natural setting.

This cautious view has as its roots in the fact that nothing is generalizable beyond the specific context where the investigation occurs.[8] Thus, because field experiments are guaranteed to help us to refine our prior about one natural setting—the causal effect that the field experiment itself estimates—they are more useful. In contrast, under this level of conservatism, laboratory experiments tell us nothing about any natural setting.

Consider the increasingly popular task of measuring social preferences. One popular tool to perform the task is a dictator game. The first dictator game experiment in economics is due to Kahneman et al. (1986). They endowed subjects with a hypothetical $20 and allowed them to dictate either an even split of $20 ($10 each) with another student or an uneven split ($18, $2), favoring themselves. Only one in four students opted for the unequal split. Numerous subsequent dictator experimental studies with real stakes replicate these results, reporting that usually more than 60% of subjects pass a positive amount of money, with the mean transfer roughly 20% of the endowment.

The common interpretation of such findings can be found in Henrich et al.'s (2004) work: "Over the past decade, research in experimental economics has emphatically falsified the textbook representation of Homo economicus, with hundreds of experiments that have suggested that people care not only about their own material payoffs but also about such things as fairness, equity, and reciprocity." Indeed, the point estimates of giving from these experiments have even been used to estimate theoretical models of social preferences (see, e.g., Fehr and Schmidt (1999)).

Under the extreme view of Proposition 20.3, such insights have limited applicability because the properties of the situation are such that we only learn about one specific situation, namely, giving in the lab. In short, our model informs us that putting subjects on an artificial margin in such a setting necessarily limits the ability to make direct inference about markets of interest.

As a point of comparison, consider a recent field measurement of social preferences from List (2006a). As discussed more fully below, one of the goals of this study was to measure the importance of reputation and social preferences in a naturally occurring setting. To explore the importance of social preferences in the field, List (2006a) carries out gift exchange natural field experiments in which buyers make price offers to sellers and, in return, sellers select the quality level of the good provided to the buyer. Higher-quality goods are costlier for sellers to produce than lower-quality goods, but are more highly valued by buyers.

The results from the AFEs in List (2006a) mirror the typical laboratory findings with other subject pools: Strong evidence consistent with social preferences was observed through a positive price and quality relationship. List (2006a) reports that similarly constructed FFEs provide identical insights. Yet, when the environment is moved to the marketplace via an NFE, where dealers are unaware that their behavior is being recorded as part of an experiment, little statistical relationship between price and quality emerges.

Viewed through the lens of Proposition 20.3, this study provides three social preference estimates that are applicable to *only* the three specific environments in which they are measured. The first estimate uses actual traders from this market in a laboratory experiment. The second uses actual traders from this market in a setting that resembles the market that they have naturally selected to participate, but one in which they know that they are being scrutinized. The third observes actual traders in a market that they have naturally selected to participate, wherein they do not know that they are being observed for scientific purposes. As such, under the extreme view of Proposition 20.3, we have at least learned about one naturally occurring setting from List's (2006a) data.

Our three propositions are summarized visually in Figure 20.2. Consider a causal triple (x, x', z) where we vary two of the dimensions of z. The space is divided into natural environments (above the dashed line) and non-natural environments (below the dashed line). One combination of (z_1, z_2) is the field experiment and one is the laboratory experiment, each of which is depicted by a spot in the figure.

Under conservative generalizability (the inner, black circles), only the field experiment yields information about natural environments. As we become less conservative and the circles expand (to the outer, gray circles), both types of experiments yield potentially disjoint information about natural environments. Thus, they become complements in the production of knowledge.

In the simpler version of the all-causes model, Falk and Heckman (2009) claim that generalizability requires an assumption of additive separability, an arbitrary assumption that is no more plausible for field experiments than it is for laboratory

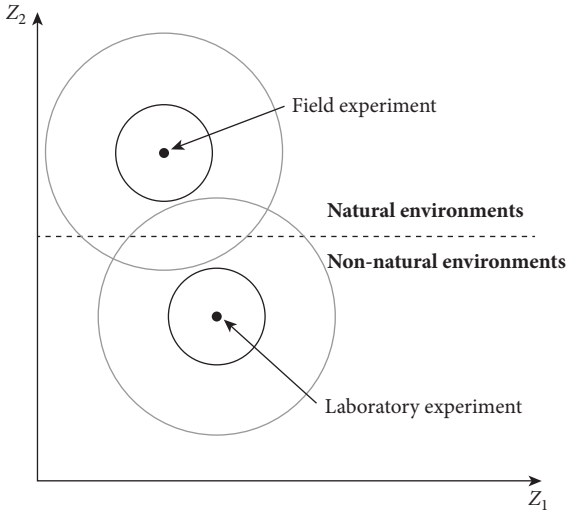


Figure 20.2. Generalizability in field and lab experiments.

experiments. However, their claim only applies for global generalizability; when generalizing locally under the assumption of continuity, additive separability is not necessary and the advantage of field experiments is particularly salient.

The kind of statistical conservatism required for zero or local generalizability is extreme, and this is because we have a highly discontinuous definition of both: Priors for certain subsets of T have to be *completely* unchanged in response to nonintersecting data. A more realistic treatment would be to include a more continuous measure of generalizability. We used highly stylized, discontinuous measures purely for expositional simplicity, akin to summarizing a hypothesis test by its conclusion (accept or reject) rather than by the p-value associated with the test statistic. The essence of our argument is unchanged by allowing generalizability to evolve into a more continuous concept.

Extending the Model: The Participation Decision

In the section entitled "Preamble: Empirical Methods," we discussed how selection impacts the measurement of treatment effects. In this section, we use our formal structure to extend the previous treatment effects discussion on the participation decision.

Consider a family of causal triples $\{g(x, x', z)\}_{z \in U_Z \subseteq S_Z}$ that an investigator wants to estimate, where z is unidimensional. Z can be thought of as a potentially observable individual-level characteristic, such as preferences or IQ. In the absence of experimental interference by the investigator, individuals learn their realization of Z and can then influence the realization of X. For simplicity, assume that at a (potentially small) cost, they can guarantee the control value, $X = x$.

We assume that it is the control rather than the treatment because usually the treatment corresponds to an intervention, whereas the control is the status quo. Conditional on the realization of Z, all remaining randomness is exogenous. Assume that at every $z \in U_Z$, a positive proportion of people are observed in each of control and treatment: $\forall z \in U_Z, 0 < Pr(X = x|Z = z) < 1$ and $0 < Pr(X = x'|Z = z) < 1$.

At this point, in principle, no experiment need be conducted. Under our highly stylized framework, the investigator can simply collect two naturally occurring observations at each value of Z (a control and a treatment) and thereby directly calculate $g\left(x, x', z\right)$. In practice, the investigator has to worry about sample sizes (the sampling issue that we abstracted away from above) and she may have a strict time limit for data collection, either of which would push her toward running an experiment where she directly and randomly manipulates the value of X.

If, after deciding to conduct an experiment, the investigator chooses to conduct it covertly (as in NFEs), then inference will proceed as normal and the desired family of causal effects will be estimated. Her ex post control over the value of X swamps individuals' ability to influence X.

On the other hand, should the investigator publicize her intention to conduct the experiment, then she has to worry about subjects exercising their *ex ante* control over X as a result of knowing about the experiment. Suppose some subset $U'_Z \subset U_Z$ decides to guarantee themselves the control value of X, meaning that the investigator cannot estimate the causal triples for this subset. The investigator has a large degree of control over X, but usually she cannot force those who, upon becoming aware of the experiment, choose not to participate. Inference for the remaining group, $U_Z \backslash U'_Z$, remains valid as before.

Consequently, she will be forced to update her priors on causal triples associated with U'_Z by extrapolating/interpolating from $U_Z \backslash U'_Z$. In practice, this will be rendered even more precarious by the possibility that Z is unobservable, meaning that the experimenter will be forced to assume that the causal triple is simply unaffected by the participation decision.[9] In the case when $U_Z = \{z_1, z_2\}$, $U'_Z = \{z_2\}$, the extrapolation bias, which we term treatment-specific selection bias, will be

$$B = g(x, x', z_2) - g(x, x', z_1).$$

Thus ironically, in a specific sense, natural field experiments afford the investigator *more* control over the environment because it allows her to bypass the participation decision. *This insight is exactly opposite to received wisdom, wherein critics argue that field experiments have less control.*

This abstract argument is illustrated above with the *Head Start* example: If parents who care more deeply about their children's outcomes are more likely to sign up for services from *Head Start*, then their children might have better outcomes in the nontreatment state than children of parents who care less deeply about their children. This orthodox selection effect is what motivates the investigator to randomize. The investigator will publicize the randomized program and solicit for

enrollment, creating the two groups $U_Z \backslash U'_Z$ (participants) and U'_Z nonparticipants. However, it might be the case that parents who believe *Head Start* will have a significant effect on their child are more likely to enroll. In that case, it would not be appropriate to generalize the effect from the $U_Z \backslash U'_Z$ group to the U'_Z group if such beliefs were actually true; the bias term B would be negative.

One potential example of this bias is randomization bias—where a direct aversion to the act of randomization is what discourages people from participating. This would be a valid concern for long-term studies where the *ex ante* uncertainty generated by randomization may lead to an expectation of adjustment costs, and hence the certainty of nonparticipation is preferred.

More generally, due to cognitive limitations, people do not take too active a role in determining natural treatment allocation in many day-to-day decisions, and so there is room for covert experimentation, for example, in how the goods are displayed in a grocery store or how a commercial looks on TV. But the very public declaration of a randomized control trial could signal the importance of a certain decision and motivate an individual to devote the cognitive resources necessary to exercise full control over participation. If you are convinced that the treatment of viewing a TV commercial is undesirable, you can just turn your TV off.

The covertness implicit in an NFE, which we are arguing is desirable, is sometimes impossible, especially in large, new programs where there is no natural, preexisting target population whose natural choices over treatment and control can be subtly manipulated by an investigator. For example, if we wanted to estimate the causal effect of introducing neighborhood watch schemes in areas with few to no neighborhood watch schemes, participation is likely to be limited in a way that interacts with the treatment effect and in a way that cannot be circumvented by covertness.

Fortunately, covertness is possible in many fields of interest, such as design of incentive schemes across many important economic domains, charitable contributions, auction design, marketing, worker compensation, organizational structure, and so on.

Advantages of Laboratory Experiments

Despite Propositions 20.1–20.3, our model strongly shows that there is a critically important advantage of laboratory experiments over field experiments. Thus far, the target space T and dataset D are exogenous. In practice, many causal triples are inestimable in field settings due to ethical/feasibility/cost reasons. For example, it is straightforward to set up a model economy in the laboratory and to manipulate randomly interest rates to gauge their effect on inflation. No such experiment is possible in a natural field experiment.

In this sense, the range of causal triples that cannot be directly estimated in a natural field experiment and that lie outside the local generalizability set of estimable causal triples is so large that in many environments, field and laboratory experiments become natural complements.[10]

Consider the case of discrimination. One would be hard-pressed to find an issue as divisive for a nation as race and civil rights. For their part, economists have produced two major theories for why discrimination exists: (i) certain populations having a general "distaste" for minorities (Becker, 1957) and ii) statistical discrimination (see, e.g., Arrow, 1972; Phelps, 1972), which is third-degree price discrimination as defined by Pigou: Marketers using observable characteristics to make statistical inference about productivity or reservation values of market agents. Natural field experiments have been importantly used to measure and disentangle the sources of discrimination (see List, 2006b for a survey).

Now consider how a laboratory experiment would be formulated. For example, if one were interested in exploring whether, and to what extent, race or gender influences the prices that buyers pay for used cars, it would be difficult to measure accurately the degree of discrimination among used car dealers who know that they are taking part in an experiment. We expect that in such cases most would agree that Propositions 20.2 or 20.3 holds.

This is not say that lab experiments cannot contribute to our understanding of important issues associated with discrimination. Quite the opposite. Consider the recent novel work of Niederle and Vesterlund (2007). They use lab experiments to investigate whether affirmative action changes the pool of entrants into a tournament. More specifically, they consider a quota system which requires that out of two winners of a tournament at least one be a woman. We suspect that this would be quite difficult to do legally in a natural field experiment. Interestingly, they report that the introduction of affirmative action will result in substantial changes in the composition of entrants.

This is just one of many studies that we could point to that serves to illustrate that, once viewed through the lens of our model, laboratory and field experiments are more likely to serve as complements as most suspect.

An aspect of laboratory experimentation that is outside of our model is another important virtue of laboratory experimentation—the ease of replication. Since replication is the cornerstone of the experimental method, it is important to discuss briefly the power of replication. Although such tracks have been covered recently by Maniadis, Tufano, and List (MTL, 2013), we parrot their discussion here because there has been a scant discussion of this important issue and it serves to highlight a key virtue of laboratory experimentation.

As Levitt and List (2009) discuss, there are at least three levels at which replication can operate. The first and most narrow of these involves taking the actual data generated by an experimental investigation and reanalyzing the data to confirm the original findings. Camerer does this with List's (2006a) data, as have dozens of other scholars. Similar to all of the other scholars, when analyzing the data Camerer finds results that are identical to what List (2006a) reported.

A second notion of replication is to run an experiment which follows a similar protocol to the first experiment to determine whether similar results can be generated using new subjects. The third and most general conception of replication

is to test the hypotheses of the original study using a new research design.[11] We show how replication and generalizability are linked using the theoretical model; our empirical example is on the second notion of replication, or whether use of the exact same protocol leads to similar empirical results. Yet, our fundamental points apply equally to the third replication concept.

Continuing with the notion that the researcher has a prior on the actual behavioral relationships, we follow MTL's model of replication. Let n represent the number of associations that are being studied in a specific field. Let π be the fraction of these associations that are actually true.[12] Let α denote the typical significance level in the field (usually $\alpha = 0.05$) and let $1 - \beta$ denote the typical power of the experimental design.[13] We are interested in the post-study probability (PSP) that the research finding is true—or, more concretely, given the empirical evidence, how sure are we that the research finding is indeed true.

This probability can be found as follows: Of the n associations, πn associations will be true and $(1 - \pi)n$ will be false. Among the true ones, $(1 - \beta)\pi n$ will be declared true relationships, while among the false associations, $\alpha(1 - \pi)n$ will be false positives, or declared true even though they are false. The PSP is simply found by dividing the number of true associations which are declared true by the number of all associations declared true:

$$\text{PSP} = (1 - \beta)\pi/((1 - \beta)\pi + \alpha(1 - \pi)) \tag{20.1}$$

It is natural to ask what factors can affect the PSP? MTL discuss three important factors that potentially affect PSP: (i) how sample sizes should affect our confidence in experimental results, (ii) how competition by independent researchers affects PSP, and (iii) how researcher biases affect PSP.

For our purposes, we can use (20.1) to determine the reliability of an experimental result. We plot the PSPs under three levels of power in Figure 20.3 (from MTL).

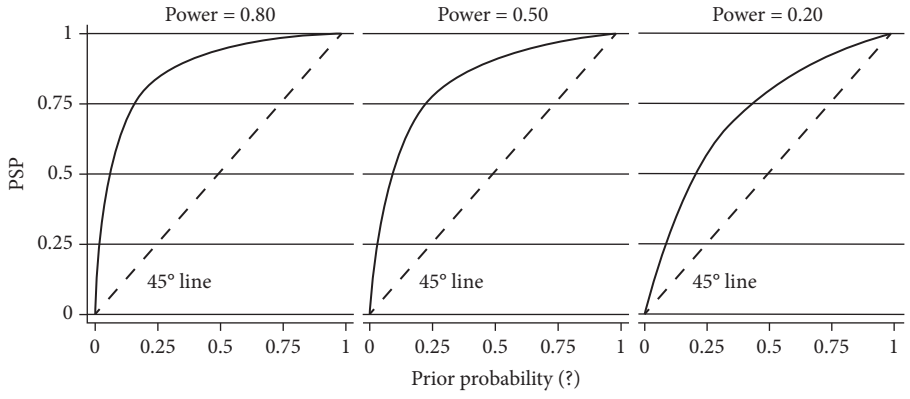


Figure 20.3. The PSP as a function of power.

Upon comparing the leftmost and rightmost panels for the case of $\pi = 0.5$, we find that the PSP in the high-power (0.80) case is nearly 20% higher than the PSP in the low-power (0.20) case. This suggests that as consumers of experimental research, we should be nearly 20% more certain that research findings from higher-powered experiments are indeed true in comparison to lower-powered experiments. What else Figure 20.3 tells us is that we should be wary of "surprise" findings of (those that arise when π values are low) from experiments because they are likely not correct findings if one considers the low PSP. In this case, they are not even true in the domain of study, much less in an environment that the researcher wishes to generalize upon.

Figure 20.3 powerfully illustrates the importance of replication for our scientific endeavors. Society can have much more confidence in our results as long as a few replications take place. The insights in Figure 20.3 are inspired by a recent study due to Moonesinghe et al. (2007), who show that even a handful of replication studies may suffice to draw the correct inference about the true association.

Of course, there are several other important virtues of laboratory experimentation, and Levitt and List (2007a, 2007b) point them out. Most importantly, LL (2007b) note that except in rare circumstances, laboratory experimentation is likely to be a useful tool for providing *qualitative* evidence, even when the generalizability of deep structural parameters is suspect.

Reflections on Commentary

The editors of this volume have asked for, and received, very important discussions of the generalizability issue. We admire the work of all the contributors. Fréchette (2015) importantly explores whether students and experts behave similarly in lab experiments. Implicitly the paper contributes to our understanding of lab treatment distributions across populations. His paper represents a great service piece, as it takes on the following invaluable question: Would one reach similar conclusions using professionals as opposed to the standard subject pool of undergraduate students?

When considering a comparison of behavior across subject pools, three important features arise: preferences, selection rules into the experiment, and behavioral effects of taking part in an experiment. To illustrate why the three features are important to consider, let's first assume that the true treatment effects $\tau_i = y_{i1} - y_{i0}$ are identical across subject pools. That is, one would obtain identical treatment effects using an NFE.

What should we find when we compare a laboratory experiment with an AFE? We very well might find significant population differences. Even in the case where the treatment effects in the student and expert population exactly overlap, selection into the experiment and scrutiny effects might cause them to look different.

Likewise, even in cases where the observed data across students and experts are statistically indistinguishable, selection or the effects of scrutiny might be masking true differences. These are potentially important aspects to consider when comparing results across groups of people, as pointed out in LL:

> One approach to investigating subject pool biases is to examine whether professionals, or other representative agents, and students behave similarly in laboratory experiments. Fehr and List (2004) examine experimentally how chief executive officers (CEOs) in Costa Rica behave in trust games and compare their behavior with that of Costa Rican students. They find that CEOs are considerably more trusting and exhibit more trustworthiness than students. These differences in behavior may mean that CEOs are more trusting in everyday life, or it may be that CEOs are more sensitive to the lab and non-anonymity effects discussed above, or that the stakes are so low for the CEOs that the sacrifice to wealth of making the moral choice is infinitesimal. (Levitt and List, 2007b, pp. 165–166)

What this means is that without the help of a selection and behavioral model, it is difficult to interpret data from simple comparisons of experts and students. Creating models that predict when behavior will be different, why it is different, and what factors exacerbate or attenuate these differences helps us explore questions such as Is behavior of students (or experts) who select into the lab representative of behavior of experts in the field? and Does a given treatment affect people in the experiment the same way it affects people in the population at large? In the end, we are after how people behave in natural settings. Yet, experimenting in important parameter spaces in the laboratory can help us tremendously in understanding behavior in natural settings. Fréchette (2015) takes us in this direction in a very welcome manner.

Kessler and Vesterlund (2015) is also a clear and thoughtful discussion of experimentation. Importantly, they concisely point out the most important misunderstood argument in LL when they note (p. 1):

> While the debate has centered on the extent to which the *quantitative* results are externally valid, we will argue that for most laboratory studies it is only relevant to ask whether the *qualitative* results are externally valid. Interestingly, among the authors on both sides of the debate there is significantly less (and possibly no) disagreement on the extent to which the qualitative results of a laboratory study are externally valid.

One of the main ideas in Levitt and List (2007a) was that the lab did not provide a good means to measure deep structural parameters. That is, LL questioned the integrity of the point estimates—for example, 63% of people have strong altruism—from laboratory exercises. The original paper discussed the various games and concluded that proper inference of results from games, such as the dictator game, had been mistaken. Studies by and large have shown the fragility of such results. During an early debate on the LL study at the ASSA meetings in Boston in January 2006, interestingly, Camerer and others on the panel strongly refuted this point.

In fact, their beliefs had recently been summarized in published work arguing that the dictator game is a useful tool to measure social preferences (Camerer and Fehr, 2004):

> Regardless of which models are most accurate, psychologically plausible, and technically useful, the important point for social scientists is that a menu of games can be used to measure social preferences, like the extent to which people weigh their monetary self-interest with the desire to reciprocate (or limit inequality), both negatively (in ultimatum games) and positively (in trust games), and with pure altruism (in dictator games).

We are happy to learn that Camerer (2015) has seemingly changed his mind and now notes that:

> It is true that many early discussions called the dictator game a way to measure "altruism", as compared to generous ultimatum offers by selfish proposers which are meant to strategically avoid rejection (e.g. Eckel and Grossman, 1996; Camerer, 2003). LL (2007a, Table 1) repeated this standard view, describing the "social preference interpretation" of dictator game giving as "altruism; fairness preferences, such as inequity aversion." The idea that dictator giving results from impure altruism, either warm glow or a preference for a social image of adhering to sharing norms, arose in the mid-1990s. The social image account was noted early on by Camerer and Thaler (1995), who used the phrase "manners." Evidence accumulated in the last few years is also consistent with aspects of warm glow and social image motive (i.e., appearing to have good "manners").

Taken together, this is a good sign that researchers are accounting for accumulated empirical evidence and changing their views accordingly. By and large, we agree with the major points of Fréchette (2015) and Kessler and Vesterlund (2015). In numerous cases, however, we disagree with Camerer's (2015) reading of our work and his interpretation of experimental work more broadly. In the interests of parsimony, we focus on the most important points of disagreement.

A first point that Camerer makes is that there is nothing inherent with features of the lab that limits their generalizability (p. 251):

> We then consider which common features of lab experiments might threaten generalizability. Are those features a necessary part of all lab experiments? Except for obtrusive observation in the lab—which is an inherent result of Federally-regulated human subjects protection in the US (though not in all countries)—the answer is "no." The special features of lab experiments which might limit generalizability can therefore be relaxed, if necessary, to more closely match particular field settings. Then we ask whether typical lab features *necessarily* undermine generalizability to all field settings in a way that cannot be corrected. They do not.

As shown above within two different theoretical frameworks, it is straightforward to see how this statement is incorrect. Indeed, we have noted several times in the literature that NFEs deliver different treatment effect estimates than the lab,

AFEs, and FFEs deliver. In a recent study, in fact, List (2011b) points this out quite directly when he notes:

> One possible source of bias in these other experimental approaches is that generally the subjects who choose to participate in the experiment are those who expect to gain the most (perhaps because they believe they are likely to get good results from the treatment). As a result, the estimated causal effect from these other experimental types, while valid, might not generalize to the target population of interest—which of course includes the subpopulation (often the majority) that did not volunteer for the experiment when offered the opportunity.
>
> Natural field experiments address this problem. Because subjects do not make a choice about whether they will participate, the treatment effect obtained from natural field experiments is, in the best-case scenario, an estimate that is both causal and broadly generalizable (in Al-Ubaydli and List, 2012, my coauthor and I offer a formal treatment). Put simply, since participants in the natural field experiment are a representative, randomly chosen, non-self-selected subset of the treatment population of interest, the causal effect obtained from this type of experiment is the average causal effect for the full population—not for a nonrandom subset that choose to participate. (List, 2011b, pp. 6–7)

Accordingly, in direct contrast to Camerer's argument, there *are* typical features of lab (and AFEs and FFEs) experiments that threaten generalizability. We hope that this study and our previous work can finally resolve that misunderstood issue.

One should then ask if there is important empirical evidence that shows the selection effect outlined above is important. LL present some evidence on both sides of the fence, and we point the reader to LL for further discussion of those studies. In our own field work, we have not found considerable evidence of selection being important in our FFEs. But, a recent innovative paper—Slonim et al. (2012)—does report strong evidence of selection being an important issue within their laboratory experiments.

Slonim et al. (2012) investigate the representativeness of participants in economics lab experiments as part of the greater research agenda of testing and interpreting the generalizability of lab experiments in economics. In particular, while a number of studies have touched upon the variance in pro-social and other behaviors within experiments by using relevant subject-level observables, there has been less emphasis on comparing the individuals that were invited to the experiment to those who eventually participated.

Slonim et al. study participation in economics experiments from a standard labor supply perspective and, using an expected utility framework, derive a set of hypotheses across a number of different subject-level dimensions. Four of the hypotheses generalize to all lab participants and are as follows (all stated relative to nonparticipants): Lab participants are predicted to have less income, more leisure time, greater interest in lab experiments, greater academic curiosity (particularly those with revealed interest in economics topics), and greater pro-sociality in terms of volunteering time.[14] The remaining hypotheses are dependent on the parameters

of the experiment itself. Specifically, the greater the uncertainty in the payoffs—participation conditional on agreeing to participate, length of the experiment, and so on, the more likely that lab participants are to be (relatively) less risk-averse. In contrast, having subjects participate by appointment only will increase the participation of risk-averse individuals relative to a "drop-in" economics experiment.

To test these hypotheses, Slonim et al. begin with a sample of 892 students in an introductory economics class which is split across 77 one-hour tutorial sections taught by 22 tutors. In the fourth week of classes, students were asked to voluntarily complete three incentivized experiments and a 20-question survey, and 96% obliged in their tutorial; after completing these, the students were presented with a random flyer advertising a lab experiment that the students could participate in anywhere between 7 and 13 days after the above tutorial section. Slonim et. al. also randomize the participation requirement over the following "appointment-type" treatments: (1) by appointment only, (2) by "drop-in," or (3) by either, that is, the subject's choice.

Looking at the main predictions of Slonim et al., there is strong support for all four of their main hypotheses in that lab participation is decreasing in spending per week (a proxy for student income; controlling for household income and hours worked), decreasing in work hours per week (used to estimate the remaining hours that could be used for leisure), increasing in interest in economics[15] and ability to make consistent decisions (both proxies for academic curiosity), and increasing in the number of times an individual volunteers a year (though not in total volunteering hours).

Additionally, they find support for the hypotheses that more risk-averse individuals choose to make an appointment when given the choice and that the propensity to save (as a proxy for patience) is predictive of participation in the lab. Using the full model and starting with a simplifying assumption that all of the above qualities are equally distributed in the population, they find overrepresentation of over 2 to 1 for the relevant income, leisure time, intellectual interest, and pro-sociality measures.

Slonim et al. also touch upon a number of suggestions for optimal design to alleviate some of the issues of nonrepresentative samples. For example, to increase the representation of higher income subjects, experimenters could increase the experimental payments or reduce the length of the experiment (for those whose high income comes not through household wealth but through hourly income). Similarly, excluding mention of economics or various social value frames can help solve the issue of overrepresentation of these types in the experimental subject pool. Alternatively, experimenters could collect the relevant observables irrespective of participation and then control for these in reporting their results. For example, when adding individuals to a mailing list to distribute information about upcoming experiments, experimenters could ask potential subjects to complete a full survey measuring the qualities that may otherwise bias results. Yet, one should still recognize that unobservables might differ across participants and nonparticipants.

The overarching point of Slonim et al.'s important contribution is that treatment effects in lab experiments can be deceiving because generalizing from experimental findings can misrepresent bias in the treatment effects. When explicitly stating an average treatment effect within an experiment, this bias only presents a problem to the extent that the particular nonrepresentativeness impacts the relevant outcome measure independently of the treatment effect. However, when looking to generalize the experimental results to economic agents in naturally occurring markets, which include participants and nonparticipants, numerous issues arise if the participation decision correlates with observable qualities of individuals that correlate with the outcome measure independently of the true underlying treatment effect.

In his own description of Slonim et al. (2012), Camerer himself concedes this point, and he suggests three remedies: using selection models, increasing the participation rate by increasing the show-up fees, and studying groups that *ex ante* have very high participation rates. We would regard these remedies as ineffective in a wide range of environments studied by economists, especially when we take into account the other advantages that we associate with natural field experiments.

Sorting of course is not merely a lab phenomenon. DellaVigna et al. (2012) highlight the importance of sorting in the field. Indeed, very rarely do economic outcomes in the field not have a participation decision associated with it, and recent field research highlights the importance of taking sorting into account. For example, Gautier and van der Kaauw (2010) examine pay-what-you-want choices by visitors to a hotel that either knew or didn't know that the hotel was pay-what-you-want before making a reservation. Visitors that knew that the hotel was pay-what-you-want *ex ante* paid significantly less than those that did not, suggesting that a different sample sorted into visiting the hotel when the information was posted.[16]

Empirical Evidence

Our final point concerns Camerer's discussion of the studies that most closely match the lab and the field. Here, beauty is certainly in the eye of the beholder because our interpretation of our own data and others' data certainly departs from Camerer's interpretation. Whether you like the List (2006a) study or not, as the reader learned from Camerer's summary of our work, when we have built a bridge between the lab and field, our own data (see, e.g., List (2004, 2006a, 2009), and the papers cited below) sometimes have good generalizability, but sometimes the lab data do not provide a good indication of behavior in the field—especially when measuring quantitative insights. We detail a few of those studies now and discuss some of the work that Camerer cites that is closely related to our own work.

We begin with the reanalysis of the data in List (2006a). Camerer is one of many scholars who have asked for the List (2006a) data (another prominent scholar who has asked for (and thoroughly analyzed) the data is Ernst Fehr; many graduate

students have also scoured the data). Each one of them has delivered replication in the first sense of the definition above: the data show what List (2006a) reported. In addition, all of them but Camerer have been satisfied with the conclusions drawn (or at least we have not heard of any dismay). To avoid a tiresome rehashing of List (2006a), we will provide a very focused response to Camerer which necessarily assumes some familiarity with List (2006a) on the reader's part. Those interested in a fuller treatment need only read the original paper (List 2006a) since we are adding nothing of substance.

A first fact (as Camerer notes) is that quality of the good delivered is much lower in the field than in the lab. Even though the same prices are paid, when dealers know that they are taking part in an experiment, they deliver higher quality. This is Camerer's first finding (p. 29), and we agree that this is what the data suggest. This is a strong pattern that is difficult to get rid of in the data, even when you try. One could stop here and the lab–field generalizability point is answered—in this setting, when a person knows that they are in an experiment, they provide higher quality. But, the data are much richer.

The crux of the other findings in the paper starts with the fact that in the field, local dealers have weakly stronger reputational concerns than do nonlocal dealers. In the field, in order of weakly increasing reputational concerns, we have periods when (1) there is no grading and no announcement of the intent to grade, (2) there is no grading but there has been an announcement of the impending arrival of grading, and (3) there is grading. There are therefore six configurations of field reputational concerns: local vs. nonlocal and no-grading versus announced grading versus grading. The key is that one can weakly rank in each dimension.

In the laboratory, local and nonlocal dealers exhibit anonymous gift giving. In multiple reputational configurations in the field, zero gift giving is observed, supporting the claim that there are zero (or weak) social preferences. Comparing behavior across reputational configurations in the field, when reputational concerns are weakly higher, anonymous gift giving is weakly higher, and in certain configurations, gift giving is very strong. This bolsters the argument that the laboratory behavior of dealers is not caused by social preferences and that it is indeed impossible to eliminate reputational concerns in the laboratory.[17]

Unsatisfied, Camerer asked for the data so that he could conduct his own investigation. He starts by pointing out that in certain configurations, gift giving in the field is stronger than anonymous gift giving in the laboratory.[18] Upon reading note 17 and List (2006a) and understanding the design and the point that is being made, it is unclear what the value of Camerer's finding is—neither List (2006a) nor us would refute this new "finding," but we find it troublesome that it is even a claim. It really has no significance whatsoever given the dimensionality of the design.

Furthermore, there is nothing to suggest that List's conclusions are driven by a lack of power in the statistical test—one just needs to look at the figures in the original study to learn about the behavioral patterns. Despite this, Camerer curiously argues that in the quest for power, he wants to discard a large portion of the data

and to focus on a small subset of the data. Camerer also claims that "it is the closest match of lab and field features," again misunderstanding the point of List (2006a).

Camerer conducts some additional tests on these data and finds that if you use some relatively unconventional tests (rather than those that List uses), some of the results switch from significance to marginal significance and, therefore, all bets are off. (As an aside, according to our attempts at replicating what Camerer did, we obtain different results for two of the three tests, though the thrust of his argument is not affected substantively.)

We describe two of the tests (Epps-Singleton and Fligner-Policello) as relatively unconventional since a Google Scholar search of papers written by Camerer where the names of these tests appear in the text yields four papers in the case of E-S and one paper in the case of F-P out of dozens and dozens of total papers published (Colin is a prolific writer). This is particularly puzzling since his supposed justification is the non-normality (in the Gaussian sense) of List's data; surely non-normality is a regular occurrence in Camerer's own data since much of it is experimental and of the same nature as our own.

Indeed, a Google Scholar search of the *American Economic Review* finds one use of the Epps-Singleton test and zero uses of the Fligner–Policello test, and analogous results are obtained for the *Quarterly Journal of Economics*, *Econometrica*, and the *Journal of Political Economy* (the first and only four journals that we checked).

Given the infrequency with which both Camerer and the literature use these tests compared to the usual array of statistical tests that we and the literature use, we regarded conducting rigorous Monte Carlo simulations to compare power as unproductive. Instead, we conjecture that most statistically significant results in economics will become marginally significant if you drop enough of the data and attempt a large enough number of statistical tests.

Camerer also chose not to comment at all about the data that he chose to throw out of the analysis. Does he think that they are useless data? For example, how does Camerer explain the finding that prior to any announcement of grading, *both* local and nonlocal dealers exhibited zero gift giving (columns 3 and 4 in Table 5 of List 2006a)? This is hardly an afterthought—it is very much the crux of List (2006a). In fact, a key to the design of List (2006a) was that it contained several outcome measures across several settings to measure behavioral change and link it to theory. Discarding hundreds of observations and focusing on a handful (as Camerer does) seems too demanding of the original data given the richness of the original design.

There is mounting evidence that the main results on generalizability found in List (2006a) are replicable and quite prominent in every setting. Winking and Mizer (2013) design laboratory and field versions of a dictator game. The goal is isolating scrutiny/knowledge of being in the experiment as the only difference between the two versions, in a similar fashion to List (2006a). The subtlety of their experimental design means that it is too long to warrant being reproduced here. However, from the perspective of Camerer's comments on List (2006a), it is reassuring to find virtually identical results: In the laboratory version of the dictator game, behavior consistent with the plethora of existing laboratory studies

was observed—that is, substantial donations. In contrast, in the field version, every single participant opted to donate nothing, choosing instead to keep all the money for themselves. Likewise, Yoeli et al. (2013) use a natural field experiment to show the observability effect (p. 1): "We show that observability triples participation in this public goods game. The effect is over four times larger than offering a $25 monetary incentive, the company's previous policy." They proceed to note that "People are substantially more cooperative when their decisions are observable and when others can respond accordingly," and they cite 20 recently published studies as further evidence.

Ultimately, the content of this exchange on List (2006a) is irrelevant to our big picture arguments about the pros and cons of field experiments. But, to complete the discussion, we now turn to a brief discussion of a few of the other studies discussed by Camerer. We are not interested in a food fight here, so we briefly comment on these strands of work. The interested reader should certainly read these other studies, the Yoeli et al. (2013) work, and the citations therein and come to an informed opinion on his/her own.

Donations, Fishermen, and Soccer

A related study (Benz and Meier, 2008) that Camerer discusses was actually published in a special issue volume that List edited on field experiments for the journal *Experimental Economics*. The Benz and Meier (2008) study permits a novel test of generalizability of results across domains, and it presents results that are quite consonant with List (2006a). Here is what List (2008) wrote about the study in his introduction, and as far as we know the data have not changed since the paper was published:

> The study takes advantage of a naturally-occurring situation at the University of Zurich, where students are asked to give money towards two social funds. The authors undertook a framed field experiment by setting up donation experiments in the lab that present students with the identical field task of giving to two social funds, but wherein they know that they are taking part in an experiment (and their endowment is provided by the experimenter).
>
> The authors are therefore able to construct a unique data set by using panel information on charitable giving by individuals both in a laboratory setting and in the field. In this respect, different from Eckel and Grossman (2008) and Rondeau and List (2008), Benz and Meier are able to change the properties of the situation without changing the population.
>
> In making comparisons across these decision environments, Benz and Meier (2008) find important evidence of positive correlation across situations, but ultimately find that giving in the lab experiment should be considered an upper bound estimate of giving in the field: subjects who have never contributed in the past to the charities gave 75 percent of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50 percent of their experimental endowment to the charities in the lab experiment.

Importantly, *subjects who have never contributed in the past to the charities gave 75% of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50% of their experimental endowment to the charities in the lab experiment*. In short, these data paint a very similar picture to what List observed amongst his sportscard dealers—the nature and extent of scrutiny affects behavior in predictable ways. This result is also found in a recent study due to Alpizar et al. (2008), who find that scrutiny is an important determinant of pro-social behavior. When done in public, charitable gifts increase by 25%, suggesting the power of scrutiny. List (2006b) provides a discussion of many more examples in this spirit.

Camerer also discusses a recent study of fishermen, due to Stoop et al. (2012). The authors present a clever assortment of experimental treatments to explore cooperation rates and why they arise in the manner in which the literature has discussed. They begin by conducting an FFE measuring cooperation among groups of recreational fishermen. They carefully construct the setting to mimic the classic voluntary contributions mechanism (VCM). Interestingly, they find little evidence of cooperation, even though the received VCM lab results represent one of the most robust performers in delivering favorable cooperation rates.

They do not stop there, however. In an effort to learn why such departures occur, they build an empirical bridge in the spirit of List (2006a, 2009) to identify the causes of the behavioral differences. They rule out the subject pool and the laboratory setting as potential causes of behavioral differences. The important variable within their environment that causes behavioral differences is the nature of the task. Importance of the nature of the task is consonant with Gneezy and List (2006) and Harrison et al. (2007). As Harrison et al. (2007) put it: Such results highlight that the controls that are typically employed in laboratory settings, such as the use of abstract tasks, could lead subjects to employ behavioral rules that differ from the ones they employ in the field. Because it is field behavior that we are interested in understanding, those controls might be a confound in themselves if they result in differences in behavior (we discuss Harrison et al. (2007) below).

This paper is important in that it provides a reinforcement of a broader point lying beneath other studies exploring behavior in quite different domains: It shows that games that have the same normal form can generate very different behavior, and it illustrates how important this can be when making laboratory and field comparisons. While Stoop et al. (2012) are able to obtain basically the same result (with modest differences) in the lab and field when the game was conducted in the same way, they find a dramatic difference in behavior when they change the nature of the task of the game, even if they change it in a way that is inconsequential in theory. This means that, for this specific case, they can isolate the exact effect leading to the lab/field difference.

Finally, Camerer also argues that the lab has advanced our understanding of game theory by testing whether subjects can play minimax in zero-sum games with a unique mixed strategy equilibrium. Minimax yields a set of stark predictions for

optimal play in such games, but decades of lab experiments (see Levitt et al. (2010) for a discussion) have found that subjects consistently deviate from optimal play, even when directly instructed to do so (Budescu and Rapoport, 1992).

Palacios-Huerta and Volij (2008) test minimax in a 2×2 and 4×4 game and find that the typical lab subjects do not play minimax, but in an AFE, subjects with experience (in particular kicking penalty shots in soccer) play just as minimax would predict. What is awkward about Camerer's choice of citing Palacios-Huerta and Volij (2008) is that this particular study finds that the standard lab experiment does not predict well when it comes to understanding behavior in AFEs or in the field. In fact, the only sample that Palacios-Huerta and Volij (2008) argue fails game-theoretic predictions are undergraduate students—that is, the only sample they find that does not play minimax is students in their lab experiment!

Nonetheless, this result should be viewed with great caution. The contribution of Palacios-Huerta and Volij (2008) to our understanding of lab or field behavior is unclear, as subsequent attempts at replication have failed (see MTL, discussed above).

In particular, Levitt et al. (2010) test whether professional soccer players, poker pros, and bridge pros can transfer their experience in the field into a lab-type setting and none do. What Levitt et al. (2010) find is that whether it is students or professionals, when they play lab games *none of them behave in accord with minimax predictions*. Thus, *if* professionals do play minimax in the field, they do not transfer that behavior to the lab games that Levitt et al. (2010) employed.

Whether actual behavior follows the predictions of minimax is still unclear, and what seems certain is that more lab and field experiments are necessary. Perhaps with big enough stakes[18] and experienced enough pros, evidence could be produced confirming minimax in a simple 2×2 game.

Open Air Markets

List (2009) uses open air markets as a natural laboratory to provide insights into the underlying operation of such markets (this discussion closely follows List (2009)). His approach reformulates and extends the problem of stability of equilibria, moving it from an abstract theoretical concept to a question about the behavior of agents in an actual marketplace. A first key result from the field experiments is the strong tendency for exchange prices to approach the prediction of competitive market equilibrium theory. Even under the most severe tests of neoclassical theory (treatments that predict highly asymmetric rents), the expected price and quantity levels are approximated in many market periods. Consonant with the above discussion on market experiments, these results suggest that in mature markets very few of the "typical" assumptions, such as Walrasian tâtonnement or centrally occurring open outcry of bids and offers, are necessary to approximate the predicted equilibrium in the field (see also List (2004)).

Yet, such markets are ripe for price manipulation. For instance, in certain cases, small numbers of sellers provide homogeneous goods that are jointly purchased from middlemen, certain barriers to entry exist, and seller communication is continual. Indeed, during the course of conducting the original tests of neoclassical theory, List learned interesting details of just such conspiracies in these markets. Armed with knowledge of a mole (confederate), he built a bridge between the lab and the field, effectively exploring the behavior of experimental subjects across sterile and rich settings. This approach has a dual benefit in that it affords an opportunity to marry the vast experimental literature on collusion in laboratory experiments (see, e.g., Holt (1995) for an excellent overview) with parallel behavior in the field.

Accordingly, he began with a general line of attack to undertake controlled experiments where factors at the heart of his conjectures were identifiable and arise endogenously. He built a bridge by beginning with data generation from a controlled laboratory study with student subjects. He proceeded to collect data using the exact same protocol with subjects from the open air market (denoted "market" subjects below) who have selected into various roles in the marketplace. He then executed a series of controlled treatments that slowly moved the environment from a tightly controlled laboratory study to a natural field experiment. By slowly moving from the typical laboratory setting to the naturally occurring environment, he is able to ascertain whether behavior differs across the lab and field domains. And, if so, he can determine the important features that drive such departures.

In this regard, comparisons of interest include observing behavior of (i) identical individuals in the lab and the field, (ii) agents drawn from the same population engaged in lab and field experiments, where the lab selection rules might be different from the way in which markets select individuals, and (iii) individuals drawn from different populations engaged in lab and field experiments.

List reports that individuals drawn from different populations show considerable signs of an ability to maintain collusive ties, even in austere situations. Yet, there are some behavioral disparities. For instance, students are influenced much more by changes in anonymity whereas marketers are influenced to a greater extent by context. For the case of agents drawn from the marketing population and placed in lab and field experimental roles, he reports only marginal evidence that selection is important. For example, when comparing cheating rates in the natural field experiment across those who agreed to participate in the lab experiments and those who refused, he finds little evidence of significant differences. However, we should highlight that this comparison is made with a sample size of 17 sellers who agreed to participate in a controlled lab or framed field treatment, along with 5 sellers who turned down the request (5 sellers were never asked) but participated (unknowingly) in the natural field treatment.

Finally, examining the behavior of the 17 individual sellers who were in experiments across the lab and the field provides insights into generalizability of results

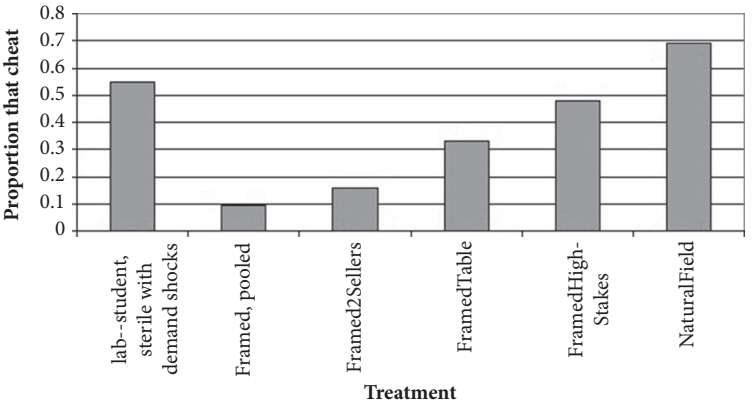


Figure 20.4. Cheating rates for executed transactions across the various treatments. Figures provide the proportion of actual transactions that were at prices below the agreed upon collusive price.

across domains. Levitt and List (2007b) argue that being part of an experiment in and of itself *might* induce certain types of behaviors. In the current case, one hypothesis is that conditional on making collusive arrangements, taking part in an experiment *might* induce sellers to more readily maintain their collusive promises. More broadly, the conditions set forth in an experimental situation might induce behavioral tendencies that are simply not observed in the field.

Using data from the 17 sellers (11 of whom were in a lab treatment, 3 of whom were in a lab and framed field treatment, and 3 of whom were in a framed field treatment), List reports a small correlation between cheating rates across the lab treatments with no context and the other environments. However, in a simple regression model, the best predictor of whether sellers will cheat in the natural field experiment is their measured cheating rate in the framed field treatments and the lab treatment with context.

Figure 20.4 highlights one aspect of List's results: rates of cheating across the experimental domains. Whereas cooperation rates are found to be quite high in the tightly controlled framed field experiments, he finds that when sellers do not know that they are part of an experiment, they cheat on the explicit agreements much more often. The level of cheating observed in the natural field experiment is larger than cheating rates observed in any of the tightly controlled framed field treatments. However, in *aggregate* the figure shows that the best predictor of cheating rates in the natural field experiment is behavior in *sterile* laboratory treatments with neutral language instructions.

Other Work

Rather than walk through every study that is published in this area, we can conclude that some lab work shows good signs of generalizability, and others do not.

We should continue to explore empirically theoretical structures such as the framework presented above to learn more about this first-order question. In this spirit, our model highlights that one area that has been largely ignored in this discussion is the fact that the typical context in which laboratory experiments are completed impose artificial restrictions on choice sets and time horizons.

One illustration of this is the work of Gneezy and List (2006), who observe work effort in two jobs where some employees are randomized into a gift treatment. The typical lab session does not exceed an hour, yet the typical labor supply decision is not made over the course of an hour. Gneezy and List find that over the course of six hours of work the gift increases worker productivity, but by the fourth hour the impact of the gift on behavior was nil. Camerer (2015) argues that the similarities of gift exchange in the lab and field during the first hour are actually evidence of the generalizability of results from the field.

Yet this more modest version of gift exchange is hardly what Fehr et al. (1993) had in mind when, in their abstract, they concluded that "These results provide, therefore, experimental support for the fair wage-effort theory of involuntary unemployment." No doubt, it is possible to do a multi-hour, real effort experiment in the laboratory—one that would cleanly examine the robustness of Fehr et al.'s results, but nobody did, confirming the original interpretation of Fehr et al. (1993).

Another illustration of how important this issue can be is found in Harrison et al. (2007). Their abstract tells the complete story:

> Does individual behavior in a laboratory setting provide a reliable indicator of behavior in a naturally occurring setting? We consider this general methodological question in the context of eliciting risk attitudes. The controls that are typically employed in laboratory settings, such as the use of abstract lotteries, could lead subjects to employ behavioral rules that differ from the ones they employ in the field. Because it is field behavior that we are interested in understanding, those controls might be a confound in themselves if they result in differences in behavior. We find that the use of artificial monetary prizes provides a reliable measure of risk attitudes when the natural counterpart outcome has minimal uncertainty, but that it can provide an unreliable measure when the natural counterpart outcome has background risk. Behavior tended to be moderately risk averse when artificial monetary prizes were used or when there was minimal uncertainty in the natural nonmonetary outcome, but subjects drawn from the same population were much more risk averse when their attitudes were elicited using the natural nonmonetary outcome that had some background risk. These results are consistent with conventional expected utility theory for the effects of background risk on attitudes to risk.

More importantly, the point that studies such as these make is that abstracting the field into the lab is not easy, and key elements of decisions made in the field, like length of work or the behavioral features that might guide decisions in the field, are missing from typical laboratory experiments.

Another example of this is Dahl and DellaVigna (2009), which compares lab and field evidence on aggression and violent movies. What previous lab

experiments failed to consider in their studies is that when someone goes to see a violent movie, they lose a few hours of opportunities to be violent. This time-use effect dominates the arousal effect frequently found in lab studies, lending to a reinterpretation of the original research findings—again, highlighting the great complementarities inherent in conducting lab and field experiments.

In sum, the totality of the evidence causes us to be quite skeptical of Camerer's other major claim:

> **Claim**: There is no replicated evidence that experimental economics lab data fail to generalize to central empirical features of field data (when the lab features are deliberately closely matched to the field features).

Just considering the evidence within our first subsection of this section: List (2006a), Bandiera et al. (2005), Benz and Meier (2008), and Alpizar et al. (2008) all find that scrutiny is an important determinant of pro-social behavior. The broad hypotheses of each study were replicated in the central features of each dataset. This caused the original lab research to not closely match the field data—from the lab, the power of social preferences would be overestimated. This is not denying the existence of such preferences, rather their import in economic settings in the field. Levitt and List (2007b) discuss further evidence. Of course, considering the evidence across the other areas discussed above reinforces the rebuttal of this claim for those areas.

In concluding, while we disagree with all of the bold claims of Camerer (2015), we agree with him fully on one count: More work needs to be done that connects outcomes in the field and the lab. There are only a handful of papers that build a bridge between the lab and the field using AFEs, FFEs, and NFEs, which is really the gold standard in mediating this discussion (see, e.g., List (2004, 2006a, 2009), and the papers cited above). Nonetheless, we have presented strong reasons to discount Camerer's reading of List (2006a), highlighting a wealth of field experiments that confirm the central contribution of List (2006a) to this debate: It can be difficult to appreciate how findings in the lab relate to observed behavior in natural settings without going to the natural settings themselves.

Epilogue

Going beyond parallelism and discussing scientifically the important issue of generalizability has been an invaluable turn for the better within experimental economics. Whereas empirical evidence is beginning to mount that helps to shed light on whether, and to what extent, received results generalize to other domains, there have been fewer theoretical advances. In this study, we put forth a theoretical model that helps frame the important features within the debate on generalizability. In doing so, it highlights the important role that field experiments should play in the discovery process.

Levitt and List (2009) discuss three distinct periods of field experiments in economics. The work of Fisher and Neyman in the 1920s and 1930s was seminal in that it helped to answer important economic questions regarding agricultural productivity while simultaneously laying the statistical groundwork relied on today. A second period of interest is the latter half of the twentieth century, during which government agencies conducted a series of large-scale social experiments. In Europe, early social experiments included electricity pricing schemes in Great Britain in the late 1960s. The first wave of such experiments in the United States began in earnest in the late 1960s and included government agency's attempts to evaluate programs by deliberate variations in agency policies. These experiments have had an important influence on policy, have generated much academic debate between structuralists and experimentalists, and anticipated the wave of recent field experiments executed in developing countries.

The third distinct period of field experimentation is the surge of field experiments in economics in the past decade or so. This most recent movement approaches field experiments by taking the tight controls of the lab to the field. Although in their infancy, this third period has produced field experiments in economics that have (1) measured key parameters to test theory and, when the theory was rejected, collected enough information to inform a new theory, (2) informed policy makers, (3) extended to both non-profit and for profit firms, and (4) been instrumental methodologically in bridging laboratory and non-experimental data. We believe going forward that field experiments will represent a strong growth industry as people begin to understand the behavioral parameters they estimate and the question they can address.

We believe that at this point the field can move beyond strong statements that *lab or field* results will always replicate. This type of reasoning seems akin to standing on the stern of the Titanic and saying she will never go down after the bow sinks below the water surface. Rather, it is now time to more fully articulate theories of generalizability and bring forward empirical evidence to test those theories. Building a bridge between the lab and the field is a good place to start. We hope that this volume moves researchers to use AFEs, FFEs, and NFEs to bridge insights gained from the lab with those gained from modeling naturally occurring data.

NOTES

We wish to thank Colin Camerer, Marco Castillo, Robert Chambers, David Eil, and Andreas Ortmann for helpful comments and discussions. Alec Brandon and David Novgorodsky provided excellent research assistance.

1. Without any evidence, we suspect that Peter Bohm was feeling similar ostracism as he presented his (seminal) challenges to laboratory experimentalists in Europe without much traction.

2. This raises the issue of informed consent. For a discussion on this, and related, issues see Levitt and List (2009) and List (2008b).

3. Various people use the term external validity. As we noted in Harrison and List (2004, p. 1033), we do not like the expression "external validity" because "what is valid in an experiment depends on the theoretical framework that is being used to draw inferences from the observed behavior in the experiment. If we have a theory that (implicitly) says that hair color does not affect behavior, then any experiment that ignores hair color is valid from the perspective of that theory. But one cannot identify what factors make an experiment valid without some priors from a theoretical framework, which is crossing into the turf of "internal validity." Note also that the "theory" we have in mind here should include the assumptions required to undertake statistical inference with the experimental data."

4. Continuity in a subset of its arguments guarantees local generalizability in a subset of dimensions.

5. This is where our allowance for non-Bayesian updating applies; a highly conservative researcher may be reluctant to update her prior if there is a large probability of the generalization being invalid.

6. We are therefore implicitly referring to NFEs (Harrison and List, 2004) when we discuss field experiments in this section, since FFEs and AFEs are not natural settings in every dimension. However, in Propositions 20.1–20.3, they will lie between NFEs and conventional laboratory experiments.

7. In this way, List's (2004) institution was more in line with Chamberlin (1948) than with Smith (1962). Since Chamberlin's original results have proven not to replicate well, we view those laboratory insights as an aberration when discussing lab results from market experiments.

8. Of course, an even more extreme view is to conclude that we can learn nothing from empirical work because of the passage of time.

9. Of course, a selection model can limit the size of the necessary leap of faith. However, unless the investigator can convincingly present a perfectly deterministic participation model, or one where residual randomness is definitively exogenous with respect to the treatment effect (neither of which is likely), then bias will remain a concern.

10. Below we give an explicit example of an important case wherein an NFE estimates an effect that is difficult (perhaps impossible) to measure in the lab.

11. This taxonomy is in agreement with the one proposed by Cartwright (1991). Hunter (2001) also defines three levels of replication, but the first level he suggests concerns the exact repetition of the original study in all dimensions, rather than the routine checking of the original study. His other two levels are largely equivalent with ours.

12. π can also be defined as the prior probability that the alternative hypothesis H_1 is actually true when performing a statistical test of the null hypothesis H_0 (see Wacholder et al. (2004)); that is, $\pi = \Pr\{H_1 \text{ is true}\}$.

13. As List et al. (2011) emphasize, power analysis is not appealing to economists. The reason is that our usual way of thinking is related to the standard regression model. This model considers the probability of observing the coefficient that we observed, if the null hypothesis is true. Power analysis explores a different question: If the alternative is true, what is the probability of the estimated coefficient lying outside the confidence interval defined when we tested our null hypothesis?

14. The authors suggest that this is distinct from pro-social behavior in terms of monetary decisions which preliminary findings suggest is correlated with wealth, introducing countervailing forces into the participation decision.

15. The authors note that given that the experiment is advertised as an "economics decision-making" task, it is difficult to rule out this potentially differential marketing as increasing this group's participation.

16. Researchers have developed techniques to deal with sorting, see the innovative study of Lazear et al. (2012).

17. Camerer also mentions the Levitt and List (2008) *Science* paper. The reader might also wish to read that study, as it is a one-page summary of the use of behavioral economics and field experiments. The point of the figure was to show that when the conditions are right, observability will affect behavior. We could have easily shown Camerer's first fact, discussed above. Or, we could have easily shown the mendacious claims data, the local versus nonlocal dealer sports card data over the third-party verification period, or the ticket stub data from both the local and nonlocal dealers. But, as evidence, we showed the data for the nonlocal dealers across the lab and field. They all showed the same behavioral pattern. Camerer thinks that we should have shown the local dealer data across the lab and field. This is quite puzzling. If he would have read the List (2006a) paper carefully, he would have known that this was the point of the entire exercise: Local dealers should show signs of gift exchange in both the lab and field, thus there should not be differences in behavior across domains because reputational concerns are driving them to engage in gift exchange (strategic reciprocity is at work)!

18. Note that this particular effect remains an open question in the economics literature. Both Smith and Walker (1993) and Camerer and Hogarth (1999) find support for the use of financial incentives over hypothetical stakes, but both suffer from opportunistic samples of existing literature. One notable exception can be found in Hertwig and Ortmann (2001) where, using a 10-year sample of studies published in the *Journal of Behavioral Decision Making*, the authors find that financial payments may improve task performance and decrease outcome variance, but also call upon using financial payments as explicitly independent variables in future experimental studies to provide a direct empirical test.

REFERENCES

Al-Ubaydli, O. and J. A. List. 2012. On the Generalizability of Experimental Results in Economics. NBER working paper series (no. 17957).

Alpizar, F., F. Carlsson, and O. Johansson-Stenman. 2008. Does context matter more for hypothetical than for actual contributions? Evidence from a natural field experiment. *Experimental Economics* **11**(3):299–314.

Arrow, K 1972. The Theory of Discrimination. In *Discrimination in Labor Markets*, eds. O. Ashenfelter and A. Rees. Princeton, NJ: Princeton University Press.

Bandiera, O., I. Barankay, and I. Rasul. 2005. Social Preferences and the Response to Incentives: Evidence from personnel data. *Quarterly Journal of Economics* **120**(3):917–962.

Becker, G. S. 1957. *The Economics of Discrimination*, 2nd edition. Chicago: University of Chicago Press.

Benz, M. and S. Meier. 2008. Do People Behave in Experiments as in the Field? – Evidence from Donations. *Experimental Economics* 11(3): 268-281.

Blundell, R. and M. Costa Dias. 2002. Alternative Approaches to Evaluation in Empirical Microeconomics. *Portuguese Economic Journal* **1**(2):91–115.

Bohm, P. 1972. Estimating Demand for Public Goods: An Experiment. *European Economic Review* **3**(2):111–130.

Budescu, D. V. and Rapoport, A. 1992. Generation of Random Series in Two-Person Strictly Competitive Games. *Journal of Experimental Psychology* **121**:352–363.

Camerer, C. 2003. *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton, NJ: Princeton University Press.

Camerer, C. 2015. The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List. In Frechette, G. R. and A. Schotter (eds). *Handbook of Experimental Economic Methodology*. Oxford University Press.

Camerer, C. F. and E. Fehr. 2004. Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, eds. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. Oxford, UK: Oxford University Press, pp. 55–95.

Camerer, C. F. and R. 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, C. F. and R. H. Thaler. 1995. Anomalies: Ultimatums, Dictators and Manners. *Journal of Economic Perspectives* **9**(2):209–219.

Card, D. and A. 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):772–793.

Cartwright, N. 1991. Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy* **23**(1):143–155.

Chamberlin, E. H. 1948. An Experimental Imperfect Market. *Journal of Political Economy* **56**(2):95–108.

Dahl, G. and S. DellaVigna. 2009. Does Movie Violence Increase Violent Crime? *Quarterly Journal of Economics* **124**(2):677–734.

DellaVigna, S., J. A. List, and U. Malmendier. 2012. Testing for Altruism and Social Pressure in Charitable Giving. *Quarterly Journal of Economics* **127**(1):1–56.

Eckel, C. C. and P. J. Grossman. 1996. Altruism in Anonymous Dictator Games. *Games and Economic Behavior* **16**:181–191.

Eckel, C. C. and P. J. Grossman. 2008. Subsidizing Charitable Contributions: A Natural Field Experiment Comparing Matching And Rebate Subsidies. *Experimental Economics* **11**(3):234–252.

Falk, J. J. and A. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* **326**(5952):535–538.

Fehr, E. and J. A. List. 2004. The Hidden Costs and Returns of Incentives—Trust and Trustworthiness Among CEOs. *Journal of the European Economic Association* **2**(5):743–771.

Fehr, E. and K. M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* **114**(3):817–818.

Ferh, E., Kirchsteiger, G. and Riedl, A. 1993. Does Fairness Prevent Market Clearing? An Experimental Investigation. *Quarterly Journal of Economics* **108**(2):437–459.

Fréchette, G. R. 2015. Laboratory Experiments: Professionals Versus Students. In *Handbook of Experimental Economic Methodology,* eds. Fréchette, G. R. and A. Schotter. Oxford University Press.

Gautier, P. A. and B. van der Kaauw. 2010. Selection in a Field Experiment with Voluntary Participation. *Journal of Applied Econometrics* doi: 10.1002/jae.1184.

Gneezy, U. and J. A. List. 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* **74**(5):1365–1384.

Harrison, G. W., and J. A. List. 2004. Field Experiments. *Journal of Economic Literature* **42**(4):1009–1055.

Harrison, G. W., J. A. List, and C. Towe. 2007. Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion. *Econometrica* **75**(2):433–458.

Heckman, A. and J. J. Falk. 2009. Lab Experiments Are a Major Source of Knowledge In The Social Sciences. *Science* **326**(5952):535–538.

Heckman, J. J. 2000. Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics* **115**(1):45–97.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford, UK: Oxford University Press.

Hertwig, R. and A. Ortmann. 2001. Experimental Practices in Economics: A Challenge for Psychologists? *Behavioral and Brain Sciences* **24**(3):383–403.

Holt, C. A. 1995. Industrial Organization: A Survey of Laboratory Research. In *The Handbook of Experimental Economics*, eds. J. Kagel and A. E. Roth. Princeton, NJ: Princeton University Press, 349-435.

Hossain, T. and J. Morgan. 2006. ...Plus shipping and Handling: Revenue (Non) equivalence in Field Experiments on eBay. *Advances in Economic Analysis and Policy* **6**(3).

Hunter, J. 2001. The Desperate Need for Replications. *Journal of Consumer Research* **28**(1):149–158.

Kahneman, D., J. L. Knetsch, and R. Thaler. 1986. Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *American Economic Review* **76**(4):728–741.

Kessler and Vesterlund. 2015. The External Validity of Experiments: The Misleading Emphasis on Quantitative Effects. In *Handbook of Experimental Economic Methodology,* eds. Fréchette, G. R. and A. Schotter. Oxford University Press.

Knight, F. H. 1921. *Risk, Uncertainty, and Profit*. New York: Cosimo.

Lazear, E., U. Malmendier, and R. Weber. 2011. Sorting in Experiments with Application to Social Preferences. *American Economic Journal: Applied Economics* **4**(1):136–163.

Levitt, S. D. and J. A. List. 2007a. Viewpoint: On the Generalizability of Lab Behaviour to the Field. *Canadian Journal of Economics* **40**(2):347–370.

Levitt, S. D. and J. A. List. 2007b. What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives* **21**(2):153–174.

Levitt, S. and J. A. List. 2008. *Science* 15 February 2008: **319**(5865):909–910.

Levitt, S.D. and J. A. List. 2009. Field experiments in economics: The past, the present, and the future. *European Economic Review* **53**(1):1–18.

Levitt, S. D., J. A. List, and D. Reiley. 2010. What Happens in the Field Stays in the Field: Professionals Do Not Play Minimax in Laboratory Experiments. *Econometrica* **78**(4):1413–1434.

List, J. A. 2003. Does Market Experience Eliminate Market Anomalies? *Quarterly Journal of Economics* **118**(1):41–71.

List, J. A. 2004. Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace. *Econometrica* **72**(2):615–625.

List, J. A. 2006a. The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy* **114**(1):1–37.

List, John A.2006b. Field Experiments: A Bridge between Lab and Naturally Occurring Data. *The B.E. Journal of Economic Analysis & Policy* **5**(2):8.

List, J. A. 2008a. Introduction to Field Experiments in Economics with Applications to the Economics of Charity. *Experimental Economics* **11**(3):203–212.

List, J. A. 2008b. Informed Consent in Social Science. *Science* **322**(5902):672.

List, J. A. 2009. The Economics of Open Air Markets. NBER Working Paper 15420.

List, J. A. 2011a. The Market for Charitable Giving. *Journal of Economic Perspectives* **25**(2):157–180.

List, J. A. 2011b. Why Economists Should Conduct field Experiments and 14 Tips for Pulling One Off. *Journal of Economic Perspectives* **25**(3):3–16.

List, J. A., S. Sadoff, and M. Wagner. 2011. So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design. *Experimental Economics* **14**(4):439–457.

Maniadis, Z., F. Tufano, and J. List. 2014. One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *American Economic Review* **104**(1): 277–90.

Moonesinghe, R., M. J. Khoury, and A. C. J. W. Janssens. 2007. Most Published Research Findings Are False—But a Little Replication Goes a Long Way. *PLoS Med* **4**(2):218–221.

Niederle, M. and L. Vesterlund. 2007. Do Women Shy Away from Competition? Do Men Compete too Much? *Quarterly Journal of Economics* **122**(3):1067–1101.

Palacios-Huerta, I. and O. Volij. 2008. Experientia Docet: Professionals Play Minimax in Laboratory Experiments. *Econometrica* **76**(1):71–115.

Phelps, E. S. 1972. The Statistical Theory of Racism and Sexism. *American Economic Review* **62**(4):659–661.

Rondeau, D. and J. A. List. 2008. Matching and Challenge Gifts to Charity: Evidence from Laboratory and Natural Field Experiments. *Experimental Economics* **11**:253–267.

Rosenbaum, P. R. and D. B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**(1):41–55.

Rosenzweig, M. R. and K. I. Wolpin. 2000. Natural "Natural Experiments" in Economics. *Journal of Economic Literature* **38**(4):827–874.

Slonim, R., C. Wang, E. Garbarino, and D. Merret. 2012. Opting-In: Participation Biases in the Lab. IZA discussion paper series (no. 6865).

Smith, V. L. 1962. An Experimental Study of Competitive Market Behavior. *Journal of Political Economy* **70**(2):111–137.

Smith, V. L. and J. M. Walker. 1993. Monetary Rewards and Decisions Cost in Experimental Economics. *Economic Inquiry* **31**(2):245–261.

Sonneman, U., C. F. Camerer, C. R. Fox, and T. Langer. 2013. How Psychological Framing Affects Economic Market Prices in the Lab and Field. *PNAS* **110**(29):11779–11784.

Stoop, J., C. N. Noussair, and D. van Soest. 2012. From the Lab to the Field: Cooperation Among Fishermen. *Journal of Political Economy* **120**(6):1027–1056.

Wacholder, S., S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. 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):434–442.

Winking, J. and N. Mizer. 2013. Natural-Field Dictator Game Shows No Altruistic Giving. *Evolution and Human Behavior* **34**(4):288–293.

Yoeli, E., M. Hoffman, D. G. Rand and M. A. Nowak. 2013. Powering up with indirect reciprocity in a large-scale field experiment. *PNAS* **110** (Supplement 2):10424–10429.

Index

Abbink, Klaus, 381–82
Abeler, Johannes, 276, 286
Adolphs, R., 191
adverse selection effect, 340, 342–43, 346–47, 354*n*3
Akerloff, George A., 151, 188, 209, 348, 352
Alatas, Vivi, 387*n*5
Alevy, Jonathan E., 377
Allais, Maurice, 14, 16, 61, 152, 161
Allais paradox, 14, 16, 152, 152*t*
all causes model, 255, 394–95, 422, 428, 436–37
Allcott, Hunt, 257
Allen, Vernon, 216
Allport, A., 167
Alpizar, F., 451, 456
"alternative-based" search (ABS), defined, 86–87
Al-Ubaydli, Omar, 6, 9, 222, 253, 288*n*17, 288*n*19, 351
Amodio, David, 5, 191, 200–201, 203
Anagol, S., 282
anchoring-and-insufficient-adjustment bias, 171–73
anchoring effects, reversal of, 172–73
Andersen, Steffen, 222, 282, 288*n*17, 316–18, 325, 333*n*23, 361
Anderson, Jon, 264–65, 284
Anderson, Lisa R., 374
Anderson, Matthew J., 386*n*2
Andreoni, James, 216–18, 221, 287*n*8, 400
Angrist, J.D., 332*n*21
anonymity, 6, 221, 223, 349, 369, 401*n*1, 453
  anonymous giving, 259, 448
  high anonymity, 223
  lab and field experiments, 215–16
  non-anonymity effects, 220, 443
  pro-sociality, 284
anonymity and lack of anonymity, lab and field experiments, 215–16
anthropology, 185–86, 194
Antonovics, Kate, 280
apparent good generalizability, imperfect design match, 277–81
Arifovic, J., 160
Armantier, Olivier, 275–76
Aronson, E., 186–87
Aronson, J.D., 190
artifactual field experiments (AFE), 423*f,* 424, 427, 436, 442, 445, 452, 456–57
  Denmark experiment, 7, 296–330
  winner's curse, 345–47, 355*n*12
artificial restrictions on choice sets, 221–22, 432, 455
Ashenfelter, O., 341
Ashraf, Nava, 226*n*3
asset integration, interpretation of risk estimates, 315–16
auction pricing, 173
auctions, 147
  common-value auctions, 107, 110, 112–16, 119
  Outer Continental Shelf (OCS) oil lease auctions, 340, 343, 345, 347, 354*n*3
  overbidding, 110, 111*t,* 112, 113*t,* 115–17, 119
  second-price sealed bid auctions, 110, 113, 119
  silent auctions, 280
  standard oral double auction (ODA), 370
  wool buyers and students, simulated progressive auction, 371–73
Aumann, Robert J., 33*n*2
authors, publication trends, 134–36, 135*f*
Autti, T., 386
Avery, C., 25

Ball, Sheryl B., 287*n*5, 361, 386*n*1
Bandiera, Oriana, 215, 412, 456
Banerjee, A.V., 35*n*15, 251, 287*n*3, 333*n*26
Banks, Jeffrey, 386*n*3
Baran, N.M., 283
Barankay, Imran, 215
Bardsley, Nicholas, 221, 254–55, 258, 287*n*8
bargaining behavior, 14, 18–20
Barr, Abigail, 281–83, 305
Barron, G., 161
Bateman, A.J., 146
Bateman, I., 278
Batson, C.D., 187, 202–3
Battalio, Raymond C., 227*n*9, 361
Baumgartner, T., 191
Bayesian Nash equilibrium, 106, 110–12, 115
Bazerman, Max, 341, 354*n*5, 355*n*6
Becker, Gary, 175, 250, 440
behavior, influencing factors, 208
behavioral investigation, professionals *vs.* students, 381–82, 382*t*

belief-based models, 106–15
accomplishments of theories, 113–14
accuracy of predications, 114
Bayesian Nash equilibrium, 106, 110–12, 115
cognitive hierarchies, 109, 114, 127*n*2
comparative static predictions, 107–9
data reconciliation, 114
estimating parameter, 107
fulfillment of underlying assumptions, 109–15
goodness-of-fit measure, 114
k-level thinking model, 106–15, 106*f*
naive play strategy, 107
out of sample behavior, 114
signals, bid classification, 111–12, 111*f,* 113*f*
subject classification, 111–12, 111*t,* 113*t*
Benartzi, S., 155
Benz, Matthias, 215, 273, 450, 456
Berg, Joyce, 170, 211, 369
Bernheim, B.Douglas, 216, 287*n*8
Bernoulli, D., 16
Berns, G.S., 192
Berrens, Robert, 216
Bewley, Truman, 352
Binmore, K., 288*n*22
biology and economics, 30, 227*n*12
Blackburn, M., 323
"blame the theory," 255
Blecherman, B., 285, 341
Block, H.D., 100
Boebel, R.B., 66
Bohara, Alok, 216
Bohm, Peter, 330*n*1, 425, 457*n*1
Bohnet, Iris, 217
Bolton, Gary E., 20, 49, 54, 61, 157, 226*n*6, 285, 288*n*18
Boly, Amadou, 275–76
Boyd, Robert T., 272, 288*n*21
Bridge pros, 410, 452
Bronars, S., 95
Brookshire, D.S., 281
Brown, M., 71
Brown, Paul, 221
Brunner, C., 22
Budescu, D.V., 452
building a research community, 14–15
Burks, Stephen V., 219
Burnham, Terence C., 217
Burns, Penny, 371, 385, 387*n*5
Buser, Thomas, 284
businessmen
students and, simulated progressive auction, 371–73
as subjects in sealed offer markets, 373–74, 373*t*
Byrnes, J.P., 128*n*11

Cacioppo, J.T., 191
Cadsby, Charles Bram, 362, 367, 385
Camerer, Colin F., 6–9, 55*n*1, 65, 76, 99, 107, 109, 153, 155, 211, 213, 226*n*3, 259, 266, 284–85, 287*n*7, 288*n*22, 309, 341, 374, 392, 395, 396, 402*n*3, 403*n*11, 432, 440, 444, 445, 447–52, 455–56, 459*nn*17–18
Cameron, Lisa, 285
Campbell, Donald, 100, 176, 187, 253, 362–63
Capen, E.C., 340, 344
Caplin, Andrew, 3, 86, 88, 100
Card, D., 423
Carlsson, Fredrik, 279
Carpenter, Jeffrey, 219, 283, 369, 385, 398
Cartwright, N., 458*n*11
Casari, Marco, 115–16, 361
Cason, T., 282
Cassing, J., 341
Castel, A.D., 190
Castillo, Marco, 397
Castronova, E., 329
Cech, Paula-Ann, 287*n*5, 361, 386*n*1
centipede game, 49–50, 49*f,* 219, 317
CEOs, trust and trustworthiness, 369
ceteris paribus observations, 208, 281, 431–32
Chabris, C.F., 282
Chamberlin, E.H., 458*n*7
charity, sharing with, 279
charity auctions, 398, 400
Charness, Gary, 6–7, 144, 157, 349–50, 355*n*6
Chen, Paul, 227*n*11
Cherry, Todd L., 219, 259, 287*nn*6–7
Chinese students and Chinese managers, 374–76, 375*t*
Chmura, T., 14
choice "mistakes," 87–88, 95–96, 98
choice process experiments. *See* enhanced choice process experiments
choice sets and time horizons, artificial restrictions, 221–22
choice task and choice rates, 155, 155*t*
Clark, A., 283, 327
Cleave, Blair L., 265
Clement, D., 250
clicking paradigm, 155–56, 155*t,* 156*t,* 160–61
closest tests of lab-field generalizability, 268–77, 449
clustering, 332*n*16
coauthorship trends, 132–33, 136–38, 136*f,* 137*f*
Coffman, L.C., 284
cognitive hierarchies, 109, 114, 127*n*2
Cohen, J., 35*n*15, 417
Cohn, A., 352
Coller, M., 325
common-value auctions, 107, 110, 112–16, 119
overbidding, 110, 111*t,* 112, 113*t,* 115–17, 119
communal fishing ponds, 274–75, 276

comparative statics
intelligent design, belief-based models, 107–9, 120
predicting, 120
professionals *vs.* students, 364, 371
verification of, 415–16
comparative *vs.* static approach, 68–69
competitive equilibrium of a market, 45–46
computable general equilibrium (CGE), 298–302
computer's rule, 108
conditional independence assumption, 423
Conlin, Michael, 212, 400
constant absolute risk aversion (CARA), 309, 312
constant relative risk aversion (CRRA), 301–2, 307, 309–10, 312, 319, 326
construct validity
neuroeconomics, 189–91, 193–94
professionals *vs.* students, 363
contests. *See* tournaments and contests
convergence approach, 146
Cook, Thomas D., 363
Cooper, David, 374, 383, 385–86
coordination games, 159, 159*t*
Cooter, Robert D., 217
Costa-Gomes, M.A., 54, 65, 108–9, 127*n*4
Costa Rica, CEO's compared to students, 220, 369, 443
Cox, J.C., 316, 399, 403*n*19, 417
Crawford, V.P., 65–66, 107–10, 127*n*4
Cronbach, L.J., 191
Croson, Rachel, 128*n*11, 218, 367, 401
cybernomics, 33*n*5

Dahl, G., 455
Dal Bo, P., 55
Darley, J.M., 202–3
Darwin, C., 146
data features and goodness-of-fit, 77–82
Davis, Douglas D., 216, 370
"Deal or No Deal" (television game show), 280
Dean, Mark, 3, 86, 88, 100
Deaton, Angus, 250
deception
polygraph test, 189–91
psychology and economics, convergence and difference, 186–88, 202–3
social psychology and economics, 168–69
decision bias, 171–72
De Cremer, D., 186
Deffenbacher, K.A., 328
Dejong, Douglas V., 373, 387*n*5
Delgado, M.R., 191
Della Vigna, Stefano, 224, 288*n*23, 447, 455
demand effect, 5–6, 260–62, 286, 287*n*12. *See also* participant bias
Denmark, artifactual field experiment and results, 7, 296–330
artifactual field experiments, role of, 321–23
asset integration, interpretation of risk estimates, 315–16
computable general equilibrium (CGE), 298–302
conclusions on, 329–30
constant absolute risk aversion (CARA), 309, 312
constant relative risk aversion (CRRA), 301–2, 307, 309–10, 312, 319, 326
discount rates, 317–21
econometrics, 297–98, 303, 318, 324–26
elicitation procedures, 305–8, 307*t,* 326–27
estimation procedures, 308–15, 310*f,* 313*f,* 315–16, 322, 326–27, 415–16
expected utility theory (EUT), 309–17, 326
external validity, 328, 333*n*25
Fechner error specification, 314–15, 333*n*26
hardnose theorist, 312, 313*f*
iterative multiple price list (iMPL), 306, 323, 326, 415
lab and field experiments
as complimentary experiments, 327–28, 417–18
contrived debate between, 323–24
disagreement with terminology, 412–16
hypothetical bias, 322–23
virtual experiments, 327–29
multiple price list (MPL), 305–6, 308
non-EUT models of risky choice and non-exponential models of discounting, 324–26
policy lotteries, 298–303, 300*f*
policy lotteries, risk aversion, 309–17, 326
randomization, 300, 333*n*26
randomization bias, 303, 323–24
risk aversion, 303–16, 326–27
risk preferences, stability of, 324
sampling procedures, steps, 304–5
de Oliveira, A., 284
descriptive accuracy and parsimony, trade-off, 151–52, 163
Dessauer, J.P., 341
Dickhaut, John W., 211
dictator game, 185–86, 188, 201–2, 216–21, 259, 279, 283–84, 449
dictator sharing, 281
double blind version of, 202, 216, 287*n*11
first experiment, 435
generally, 210*t*
money sharing, lab–field generalizability, 276, 444
Dietz, T., 14
dining game, 215
discount rates, 317–21
doctors, designing labor markets for, 22–29
ecological validity, 29

doctors, designing labor markets for, (*continued*)
gastroenterology fellows, redesign of labor market for, 26–29
gastroenterology markets, 23*t,* 25–26
new medical graduates, 22–26, 23*t*
theory of stable matchings, 22–23
donations to student funds, 273–74
dopamine system, reward prediction error, 88, 100
Doty, Richard L., 219
Douglas, R.W., 341
dropouts, testing theories in the lab, 74–75
Dubner, S., 259
Duflo, Esther, 35*n*15, 251, 287*n*3, 333*n*26, 417
Dufwenberg, Martin, 4, 330*n*1
Dupas, P., 417
Dyer, Douglas, 344–45, 347, 361, 378, 380, 403*n*10, 409, 413

Easterly, W., 35*n*15
Eckel, Catherine C., 128*n*11, 220, 265, 450
ecological validity, 29, 31, 187–88, 201, 344, 362, 364
econometrics
Denmark, artifactual field experiment and results, 297–98, 303, 318, 324–26
"theory, experimental design and econometrics are complementary," 412–16
effort, defined, 349
effort level, 74, 222, 349–50
Einav, L., 137
elicitation procedures, 305–8, 307*t,* 323, 326–27, 415
elimination design, 118–20, 127
Ellison, G., 137
Ellsberg, D., 16, 61
emotions, 174
empirical evidence, 447–56
donations, fishermen, and soccer, 450–52
monetary incentives, 450
open air markets, 272–73, 452–54, 454*f*
other work, 454–56
unique mixed strategy equilibrium, 451
voluntary contribution mechanism (VCM), 451
zero-sum game, 451
empirical labor economists, 407
empirical methods, 422–27
conditional independence assumption, 423
field experiment bridge, 423, 423*f,* 424
minimum wage laws, quantity of labor supplied, 423
modeling approaches, 423–24
propensity score matching, 423, 423*t*
endocrinology and economics, 30, 34*n*11
endowment effect, 277–78, 288*n*23
Engell, A.D., 191
Engelmann, D., 288*n*23
Engers, M., 398
enhanced choice process experiments, 3, 86–101
"alternative-based" search (ABS), defined, 86–87
choice alone, 97–99
choice "mistakes," 87–88, 95–96, 98
choice process data, 100–101
advantages, 99
theory, 88–93
choice process experiment, 93–97
conclusions on, 101
dopamine system, reward prediction error, 88, 100
methodology, 99–101
neuroeconomic research field, 99–100
"pure choice," 97–98
"reservation-based" search (RBS), defined, 87
satisficing model, 87, 91, 96–97
search order, 97
sequential search, 95
Ensminger, J.E., 287*n*14
epilogue, 456–57
epsilon equilibrium, 47–48, 50, 53*f*
equilibrium theory, 44–46
centipede game, 49–50, 49*f*
competitive equilibrium of a market, 45–46
conclusions on, 55, 82–83
epsilon equilibrium, 47–48, 50, 53*f*
experiments, 46–55
history of bids, 45, 46*f*
matching pennies game, 52–53, 53*f,* 64*t*
Nash equilibrium, 44–55, 53*f*
quantal response equilibrium, 44, 50–53, 53*f,* 55, 62
quantal response equilibrium *vs.* Nash equilibrium theory, 63–65, 64*t*
self-confirming equilibrium, 48–49
selfish preferences theory, 44, 46–47, 52–53
ultimatum game, 46–49, 47*t*
voting experiment, 44–45, 45*f,* 50–52, 51*f*
Erev, Ido, 5, 17, 20, 76–77, 77, 121, 155, 160–63
Ericsson, K., 99
Ert, E., 17
event partitions in prediction markets, 278–79
executives and students
winner's curse, field experiments, 344–45
expected utility theory (EUT)
non-EUT models of risky choice and non-exponential models of discounting, 324–26
policy lotteries, risk aversion, 309–17, 326
role of theory in experiments, 61–62
experimental economists and laboratories, statistics, 14
experimental games, summary, 210*t*–211*t*
experimental results in economics, generalizability, 9, 420–57
all causes model, 422, 428, 436–37
artifactual field experiment (AFE), 423*f,* 424, 427, 436, 442, 445, 452, 456–57
artificial restrictions on choice sets, 432

ceteris paribus observations, 431–32
context, embedded decision, 263–64
empirical evidence, 447–56
empirical methods, 422–27
epilogue, 456–57
formalizing generalizability, 427–42
framed field experiment (FFE), 423*f,* 424–25, 427, 436, 445, 451, 456
Head Start firms, example, 426–27, 438–39
identification, estimation of treatment effects, 422–23, 426–27, 429–30
lab participants economic experiments, representativeness, 445–47
moral or ethical considerations, 258–60
natural field experiment (NFE), 423*f,* 425, 427, 436, 438–39, 442, 444, 456–57
naturally occurring data, 423–25, 423*f,* 429, 457
nature and extent of scrutiny of one's actions by others, 260–63
new experiments, remedy for concern about older experiments, 267
parameters estimated, 425–27
potential threats to generalizability, 257–67
qualitative results, external validity of, 443
randomization
bias, 427, 439
empirical methods, 426–27, 430, 438
participation decision, 438–39
real world experiments and, 432, 434
replication, 422, 440–42, 441*f,* 448–49, 452, 458*n*11
reputation, consequences and concerns, 436
selection bias, 266, 426–27, 438, 443, 445–47
self-selection of the individuals making the decisions, 264–66
stakes of the game, 266–67
experimenter maxims, 4, 141–44
journal publication, submission for, 143
producing good experiments and writing results
formula, 141–42
what not to do, 142–43
experiment simplicity, 147
explanatory and predictive theory, distinctions, 62–67
external validity, 251–56, 285–86
Denmark, artifactual field experiment and results, 328, 333*n*25
lab experiments, misleading emphasis on quantitative effects, 8, 391–401, 402*n*5
all causes model, 394–95
anonymity, 401*n*1
charity auctions, 398, 400
conclusions on, 399–401
lab and field results as complementary, 8, 391–401, 402*n*5
naturally occurring data, 392
other regarding preferences, 400
parallelism, 396
promise of generalizability, 395–96
quantitative results *vs.* qualitative external validity, 393–400, 402*n*4
replication, 402*n*2
scientific view, 395
sports card trading, 397, 401*n*2
world outside the lab, 396–99
origination of term, 253–54
qualitative results, 443
social psychology and economics, 171, 175–76, 178
testing theories in the lab, 71
winner's curse, 340, 343, 344–47

Falk, Armin, 197–98, 199*n*4, 253, 255–56, 261, 267, 282, 286, 394, 396, 402*n*5, 432, 436
Fechner error specification, 314–15, 333*n*26
Feeny, David, 387*n*4
Fehr, Ernst, 20, 49, 54, 61, 157, 182, 211, 213, 218–20, 283, 349, 369, 443–44, 447–48, 455
Feng, L., 288*n*23
Fessler, Daniel M.T., 216
Feyerabend, P., 59
field experiment bridge, 423, 423*f,* 424
field experiments. *See* lab-field experiments
financial market professionals, information cascades, 377–78, 378*t*
Fiore, Stephen M., 322, 328–29
Fischbacher, Urs, 218–19
fishing
communal fishing ponds, 274–75, 276
Toyama Bay, Japanese fishing community, 283, 369–70
flea markets, 272–73, 452–54, 454*f*
Flinn, Chris J., 71, 83*n*4
Flood, M.M., 15
fMRI (functional magnetic resonance imaging), 29, 188, 190–91, 192, 193
formalizing generalizability, 427–42
dictator game, 444
field experiments advantages, theory and propositions, 431–37, 437*f*
lab experiments, advantages, 439–42, 441*f*
participation decision, 437–39
qualitative results, 392–93, 396–98, 399–400, 402*n*4, 443
quantitative results, 392–93, 395, 400, 402*n*4, 443
setup, 428–30
types of generalizability, 430–31
Fouraker, Lawrence E., 365, 387*n*10
framed field experiment (FFE), 423*f,* 424–25, 427, 436, 445, 451, 456
Franzen, Axel, 281

Fréchette, Guillaume, 8, 83*n*5, 202, 287*n*5, 424, 442–44
Frederick, S., 332*n*20
free-riding or socially efficient equilibrium, 367–68, 368*f*
Friedman, Milton, 16, 34*n*12, 62–63, 66–67, 82, 128*n*9
Frykblom, Peter, 219
Fudenberg, Drew, 48–49, 54–55
fungibility of cash and in-kind subsidies, 276

Gächter, Simon, 211, 213
Gale, D., 23
Galinsky, Adam, 172
game shows, 223, 254
  "Deal or No Deal" (television game show), 280
  discrimination, 223
  field data, 254
  "The Weakest Link" (television game show), 223, 280–81
gaming against managers in incentive systems
  information signals, 374–76, 375*t*
Gans, J.S., 137
Garrat, R., 347
gastroenterology markets, 23*t,* 25–29
gender and competition, 146
  tournament selection example, 122–26, 125*f,* 283, 440
generalizability
  experimental results (*See* experimental results in economics, generalizability)
  measuring social preferences (*See* measuring social preferences, lab experiments)
generalization with substantial design and subject differences, 281–85
generalizing the findings of lab experiments to actual markets, 222–24
Genesove, D., 341
Georganas, S., 109
gift exchange, 210*t,* 214, 218–22, 269–70, 269*t,* 283
  anonymity, 349
  conclusions on, 353–54
  effort, defined, 349
  effort level, 349–50
  experimental labor markets, 7–8, 348–53, 356*n*26, 409–12
  higher wages, 348–49
  lab experiments, measuring social preferences, 214, 218–22, 269–70, 269*t,* 283
  labor relations in firms, relevance and background, 348–52
  measurement issues, 352
  positive and negative reciprocity, 349–53, 355*n*22
  reputation, consequences an concerns, 349
  statistical specification, 356*n*26
  sticky downward wages, 352–53
  wage offers, 411–12
gift treatment, 222, 455
Gigerenzer, G., 327
Gilovich, T., 171–72, 172
Glaeser, Edward, 224, 226*n*2
Glimcher, P., 188, 327
Gneezy, Uri, 4, 30, 122, 128*n*11, 215, 222, 227*n*11, 287*n*2, 326, 350–51, 356*nn*25–26, 361, 412, 451, 455
Gode, Dhananjay K., 370
Goeree, Jacob K., 52, 63–64, 398
Gollwitzer, Peter M., 192
good experimental design. *See* intelligent design
goodness-of-fit measure
  belief-based models, 114
  testing theories in the lab, 72–73, 77–82
Good Samaritan parable, 203
Grampp, William, 226*n*3
Greiner, Ben, 5
Grether, D.M., 21
Grosskopf, B., 119
Grossman, Phillip, 128*n*11, 220, 265, 450
Gul, F., 49, 54
Güth, Werner, 20, 61, 83*n*2, 211

Haley, Kevin J., 216
Hannan, R.L., 350, 355*n*22, 356*n*28
Harbaugh, William T., 17, 227*n*10, 361
Harbring, C., 83*n*6
hardnose theorist, 312, 313*f*
hard science, 189–91
Harris, S.J., 216
Harrison, Glenn W., 7, 9, 128*n*9, 186, 299–300, 302, 304–6, 311–12, 315, 317–18, 321–23, 325, 327–28, 330*n*1, 331*n*6, 333*nn*24–25, 334*n*29, 345–47, 355*n*12, 407, 412–15, 424, 451, 455, 458*n*3, 458*n*6
Harsanyi, J.C., 33*n*2, 54
Hartshorne, Hugh, 215
Haruvy, E., 25, 215
Head Start firms, example, 426–27, 438–39
Heckman, James J., 71, 197–98, 253, 255–56, 261, 267, 286, 394, 396, 402*n*5, 428, 432, 436
hedonic pricing method (HPM), 329
Heller, D., 65
Hendricks, K., 355*n*8
Hennig-Schmidt, Heike, 222, 356*n*25
Henrich, Joseph, 185, 217–18, 264, 287*n*5, 435
heritability of economics, 30
Hertwig, Ralph, 161, 217, 459*n*18
Hey, J.D., 309, 314, 332*n*15
Ho, T.-H., 66, 76, 309
Hoffman, Elizabeth, 216, 226*n*6, 287*n*11
Hoffman, M.H., 264
Hoffman, R.R., 328
Hogarth, Robin M., 266, 459*n*18

Hollard, G., 288*n*23
Holt, Charles A., 52, 64, 199*n*3, 207, 216, 305, 311, 312, 314, 319, 370
Hossain, Tanjim, 425
hot hand in basketball, 171–72
Houtman, M., 95
Hunter, J., 458*n*11
hypothesis testing, 105, 115–21, 127
  comparing predictive power of theories, 120–21
  elimination design, 118–20, 127
  Old Testament example, 117–18
  testing for selection, 115–16
  two-way design, 116–18, 127
hypothetical bias, 322–23

Ichino, A., 282
identification, estimation of treatment effects, 422–23, 426–27, 429–30
implications for experiments, 213
incentive structure, 160
individual choice behavior, 16–18, 19, 30, 216
individual's choices, testing theories, 77, 79*t*
influencing factors of behavior, 208
information cascades, 377–78, 378*t*
information signals, 374–78
intelligent design, 3–4, 104–27
  belief-based models, 106–15
  channel driven results, 121
  common-value auctions, 107, 110, 112–16, 119
  comparative statics, 107–9, 120
  competition and gender, tournament selection example, 122–26, 125*f*, 283, 440
  conclusions on, 127
  hypothesis testing, 105, 115–21, 127
  treatment-driven experiments, 121–26
  two-player games, 108, 119
    common value auction, 110
    second-price sealed bid auctions, 110, 113, 119
    unique mixed strategy equilibrium, 121
    zero-sum games, 121
investigation neutrality, 253
Iriberri, N., 65–66, 107, 110
Irlenbusch, B., 83*n*6
Isaac, R.Mark, 280
iterative multiple price list (iMPL), 306, 323, 326, 415
Ivanov, A., 110, 119

James, William, 167
Jamison, J., 169
Japanese fishing community, 283, 369–70
Jastrow, J., 333*n*26
Jevons, William Stanley, 34*n*10
joint estimation, 319–21, 320*t*
journal publication, 4, 132–39
  authors, publication trends, 134–36, 135*f*
  coauthorship trends, 132–33, 136–38, 136*f*, 137*f*
  conclusions on, 138–39
  data, 133
  experimenter maxims, 4, 141–44
  microeconomics faculty, top 15 school rankings, 139*n*2
  papers, publication trends, 133–34, 133*t*
  publication trends, 133–36
  submission of research, experimenter maxims, 143
    request to editors and referees, 143–44
    sealed-envelope-submission strategy, 143
  theorists and experimentalists, generally, 132–33
  theorists dabbling in experiments, 138, 138*f*, 139*n*8

Kagel, John H., 7, 9, 22, 24, 27, 117, 128*n*9, 227*n*9, 285, 288*n*20, 341–43, 345, 347, 350, 355*n*6, 361, 363, 365, 379, 395, 407, 409, 410–13
Kahneman, Daniel, 14, 16–17, 17, 29, 55, 61, 155–56, 161, 171–72, 211, 327, 435
Karlan, Dean, 283
Katok, Elena, 226*n*6, 399, 403*n*19, 417
Keller, L.R., 332*n*20
Kelley, Hugh H., 170, 178
Kerkvliet, Joe, 216
Kessler, Judd, 8, 279, 398–99, 402*n*5, 443–44
Khasi speakers (India), 288*n*18
Kimmel, M., 255
King, Ronald R., 387*n*4
Kirchsteiger, Georg, 211
k-level theory, 65–66, 106–14, 106*f*, 127*n*2, 127*nn*4–5
Knetsch, Jack L., 211
Knight, F.H., 420
Knutson, B., 192
Kranton, Rachel E., 209
Krause, Kate, 17, 227*n*10
Kremer, Michael, 417
Krueger, Alan B., 227*n*11, 423
Krupka, Erin L., 287*n*8
Ku, Gillian, 172
Kube, Sebastian, 222, 351
Kwasnica, A.M., 399, 403*n*19, 417

lab experiments
  external validity, misleading emphasis on quantitative effects (*See* external validity)
  field experiments and (*See* lab-field experiments)
  social preferences, measuring (*See* measuring social preferences, lab experiments)
lab-field experiments
  apparent good generalizability, imperfect design match, 277–81
  as complimentary, 327–28, 412–18, 417–18
  contrived debate between, 323–24

lab-field experiments (*continued*)
Denmark, artifactual field experiment and results, 7, 296–330
empirical and experimental economics, 9, 407–18
field experiments, lack of control, 408
generalization with substantial design and subject differences, 281–85
gift exchange, experimental labor markets, 7–8, 348–53, 356*n*26, 409–12
hypothetical bias, 322–23
measuring social preferences (*See* measuring social preferences, lab experiments)
virtual experiments, 327–29
winner's curse, 7–8, 340–47, 409–12
labor markets, wages and. *See* gift exchange, experimental labor markets
labor relations in firms, relevance and background, 348–52
lab participants economic experiments, representativeness, 445–47
Lamba, S., 284
Lambdin, C.G., 261
Landry, Craig E., 284
LaPiere, R.T., 183
Lau, Morten, 7, 9, 407
Laury, Susan K., 215, 226*n*6, 305, 311–12, 314, 319
Lazear, Edward P., 83*n*6, 221, 459*n*16
Ledyard, John O., 21, 211, 226*n*4, 370
Lee, Darin, 227*n*11
Leibbrandt, A., 274, 283
level k theory. *See* k-level theory
Levin, Dan, 117, 285, 341–42, 355*n*6, 379, 395, 413
Levine, David K., 3, 44, 48–50, 54–55, 83*n*1
Levitt, Steven, 2, 6, 8–9, 31, 197–98, 208, 223, 250, 252–55, 258–59, 262–65, 267, 269–72, 274, 287*n*9, 327, 392–94, 396, 398, 402*nn*5–6, 403*n*14, 410, 414, 421–22, 440, 442–44, 452, 454, 456–57, 459*n*17
Liday, Steven G., 227*n*10
lie detector (polygraph), 189–91
List, John A., 2, 6, 8, 9, 31, 197–98, 199*n*4, 208, 214, 216, 220–22, 227*n*11, 250–55, 258–59, 262–65, 267–72, 277–78, 283, 287*nn*8–9, 288*n*17, 288*n*19, 302, 315, 321, 323, 327–28, 333*n*25, 345–47, 350–51, 355*n*12, 355*n*16, 356*nn*25–26, 369, 392, 393–94, 396–98, 401*n*2, 402*n*5, 402*n*6, 403*n*14, 412–15, 420–22, 424–25, 433–34, 436, 440, 442–45, 447–57, 458*n*3, 458*nn*6–7, 459*n*17
Liu, E.M., 282
lobbying experiment, information signals, 376–77, 376*t*
Loewenstein, George, 222, 226*n*3, 226*n*5, 402*n*4
long-run decisions, 222, 227*n*12
consequences, 329
lotteries
policy lotteries, Denmark, 298–330
Swedish lotteries, 279
Lount, R.B., 174
Lucas critique, 226*n*7
Lucking-Reiley, D., 355*n*16, 399, 403*n*19, 417
Lusk, Jayson, 282
lying. *See* deception
Lynn, Michael, 212

Mace, R., 284
Maggi, G., 282
Maks, J.A.H., 95
Malmendier, Ulrike, 221, 224
Malouf, M.K., 18–19
managers
coordination failure, experienced managers, 383–84, 383*t,* 384*t*
gaming against managers in incentive systems, 374–76, 375*t*
Maniadis, Z., 440
Maréchal, Michel André, 222
market design
cybernomics, 33*n*5
designing labor markets for doctors, 22–29
first attempts, 21
as ongoing process, 22
market experiments, 370–73
experience and decision making, 371–73, 372*t*
professionals *vs.* students, 370–73
standard oral double auction (ODA), 370
Marklein, F., 276
Marks, Melanie, 367
Marr, D., 327
Marschak, J., 100
Martin, Daniel, 86, 288*n*22
Mas, Alexandre, 227*n*11
Masatlioglu, Y., 102*n*5
matching pennies game, 52–53, 53*f,* 64*t,* 274
mathematical formalism, success of, 146
mature industrial personnel and college students, comparison, 365–67
May, Mark A., 215
Maynes, Elizabeth, 362, 367, 385
McAfee, Preston, 348, 412
McCabe, D.P., 190
McCabe, Kevin A., 211, 216–17, 226*n*6
McGrath, Joseph, 175, 177–78, 178, 198
McKelvey, R.D., 49–50
McKinney, McKinney, C.N., 26
McManus, B., 398
McMillan, J., 348
Mead, W.J., 343, 355*n*8
measuring social preferences, lab experiments, 207–25, 254–67
anonymity and lack of anonymity, lab and field experiments, 215–16
centipede game, 219, 317

choice sets and time horizons, artificial restrictions, 221–22
conclusions on, 224–25, 285–86
contextual factors, 217–19
critical reply to Levitt and List, 249–86
demand effect, 5–6, 260–62, 286, 287*n*12
effort level, 222
external validity, 251–56, 285–86
game shows, 223, 254
generalizability, critical reply to Levitt and List, 254–67
apparent good generalizability with imperfect lab–field design match, 277–81
closest tests of lab–field generalizability, examples, 268–77
comparing methods for empirical inference, 255–57
context, embedded decision, 263–64
generalization with substantial design and subject differences, 281–85
moral or ethical considerations, 258–60
nature and extent of scrutiny of one's actions by others, 260–63
new experiments, best remedy for concern about generalizability of older experiments, 267
potential threats to generalizability, 257–67
scientific view, 252–54
selection bias, 266, 426–27, 438, 443, 445–47
self-selection of the individuals making the decisions, 264–66
stakes of the game, 266–67
generalizing the findings of lab experiments to actual markets, 222–24
gift exchange, 214, 218–22, 269–70, 269*t,* 283
implications for experiments, 213
influencing factors of behavior, 208
moral choice, 208, 209–12, 219, 258–60
naturally occurring data, 208–9, 227*n*11, 264
prisoner's dilemma, 210*t,* 211*t,* 217–18, 223
replication, 288*n*19
scrutiny, 213–15, 224
stakes, morality and wealth components, 219
study participants, selection, 219–21
summary of experimental games, 210*t*–211*t*
ultimatum game, 210*t,* 217–19
utility maximization model, 209, 211–12
"The Weakest Link" (television game show), 223
Meehl, P.E., 191
Meier, Stephan, 215, 273, 282, 450, 456
Merlo, A., 128*n*9
Mestelman, Stuart, 387*n*4
Miguel, E., 417
Milgram, S., 203
Milgrom, Paul R., 21, 33*n*5, 343
Miller, John, 218, 305
Mischel, Walter, 215
Mitzkewitz, M., 157
Mizer, Nicholas, 449
Moldovanu, B., 73
Monderer, D., 54
monetary incentives, 169, 203, 450
money sharing (dictator game). *See* dictator game
Montague, P.R., 192
Monte Carlo analysis, 301, 449
Montmarquette, Claude, 387*n*6
Moonesinghe, R., 442
moral choice, 208, 209–12, 219, 220, 258–60, 443
lab experiments, measuring social preferences, 209–12, 219, 220
model of utility with wealth and morality, 209–12
stakes, morality and wealth components, 219
threats to generalizability, 258–60
Morgan, J., 264, 425
Morgenstern, O., 61, 152
Mueller, W., 73–74
Mullainathan, S., 257
multiple price list (MPL), 305–6, 308
iterative multiple price list (iMPL), 306, 323, 326, 415
Munger, K., 216
Murnighan, Keith, 5, 19, 166–68, 174
Murningham, J.K., 198
Myers, Caitlin Knowles, 283

Nagel, R., 106–7, 119, 157
naive play strategy, 107
Nakajima, D., 102*n*5
Nash, John, 33*n*2
Nash equilibrium, 18–20, 44–55, 53*f,* 63–65
Bayesian Nash equilibrium, 106, 110–12, 115
new Nash equilibrium, 53*f*
original Nash equilibrium, 53*f*
quantal response theory *vs.,* 63–65, 64*t*
Rand Corporation, 15
risk-neutral Nash equilibrium (RNNE), 116–17, 342, 344
natural field experiment (NFE), 423*f,* 425, 427, 436, 438–39, 442, 444, 456–57
naturally occurring data, 208–9, 227*n*11, 264, 392, 423–25, 423*f,* 429, 457
Neelin, J., 20
neuroeconomic research field, 99–100
neuroeconomics, 188–94
construct validity, 189–91, 193–94
hard science, 189–91
polygraph test, 189–91
"true" feelings inferred from physiology, 189–91
trust, 182, 190–93
value, 192–93

neuroimaging and behavioral economics, 191–94
fMRI (functional magnetic resonance imaging), 188, 190–91, 192, 193
new areas of research, 29–30
economics and biology (E&B), 30
economics and psychology (E&P), 29–30
endocrinology and economics, 30, 34*n*11
new medical graduates, 22–26, 23*t*
gastroenterology markets, 23*t,* 25–26
theory of stable matchings, 22–23
Niederle, Muriel, 3–4, 22, 26–30, 122, 126, 361, 440
Nisbett, Richard E., 215
Nobel Prize history, parallel of experimental economics growth, 15
Nordhaus, W., 249
Nosenzo, Daniele, 286
nurses *vs.* economics and business students, 367–68, 368*f*
Nyarko, Y., 77, 80–82

obtrusive-unobtrusive research operations
social psychology and economics, 175, 176*f,* 198
Ochs, Jack, 20
Ockenfels, Axel, 20, 49, 54, 61, 157, 288*n*18
O'Donoghue, Ted, 212
Offerman, Theo, 6
Old Testament, 117–18
Onderstal, S., 284, 398
1-800 critique, 5, 151–63
Allais paradox, 152, 152*t*
cards, 155*t,* 156, 161
choice task and choice rates, 155, 155*t*
clicking paradigm, 155–56, 155*t,* 156*t,* 160–61
coordination games, 159, 159*t*
counter-to-counterexamples, shortcomings, 5, 154–62
incentive structure, 160
individual decision making experience, weighting rare events, 155–56, 155*t,* 156*t*
prisoner's dilemma, 152, 153*t*
prospect theory, 152, 155, 162
quantitative predictions, 159–62
rare events, weighting, 152, 154, 155–56, 155*t,* 156*t,* 160, 162
social conflict experience, 156–59, 157*t,* 158*f,* 159*t*
stag hunt game, 157–59, 157*t,* 158*f*
ultimatum game, 152, 153*t,* 157
O'Neil, Barry, 66
O'Neill, 380
one-shot games, 211, 218
decisions, 161, 216–17
experimental designs, 218
open air markets (flea markets), 272–73, 452–54, 454*f*
optimism bias, 132
option pricing by students and professional traders, 381–82, 382*t*
oral double auction (ODA), 370
Orme, C., 309, 314, 332*n*15
Orne, Martin T., 213, 262, 287*n*12
Ortmann, Andreas, 217, 327, 459*n*18
Orzen, H., 398
Östling, Robert, 279
Ostrom, Elinor, 14
other regarding preferences
anonymity, 369
bargaining behavior, 365–67
external validity of lab experiments, misleading emphasis on quantitative effects, 400
hidden costs and returns of incentives, 369
professionals *vs.* students, 365–70
psychology research and economics research, 152, 157
socially efficient or free-riding equilibrium, 367–68, 368*f*
social preferences, productivity, 369–70
voluntary contribution mechanism game (VCM), 360–70
Outer Continental Shelf (OCS) oil lease auctions, 340, 343, 345, 347, 354*n*3
out of sample behavior, 114
overbidding
common-value auctions, 110, 111*t,* 112, 113*t,* 115–17, 119
winner's curse, 342–43
Owens, M.F., 350

Packard, T., 282, 305
Palacios-Huerta, Ignacio, 274, 380, 385, 387*n*12, 410, 452
Palfrey, Tom R., 44, 49, 50
papers, publication trends, 133–34, 133*t*
Papke, L.E., 332*n*15
parallelism, 252–53, 263, 396, 453, 456
parameters estimated, 425–27
Parco, James E., 219
parsimony and descriptive accuracy, trade-off, 151–52, 163
participant bias, 185–88, 201–3, 266
Pashler, H., 72
peer effects, 282
Peirce, A.H., 333*n*26
Peranson, E., 22, 26
perfect equilibrium, 2, 14, 19–20, 46
subgame perfect equilibrium, 46, 83*n*2, 157, 370
Pesendorfer, W., 51, 54, 153–54
Pesendorfer's assertion, 153–54
physical sciences, 168, 207–8
Pierce, A.H., 213, 262
Pigou, Arthur C., 440
Pischke, J.S., 332*n*21
Plassmann, H., 192
Platt, John, 171
Plott, Charles, 21, 45, 147, 255

Pointner, Sonja, 281
poker players, 274, 410, 452
Poldrack, R.A., 192
policy lotteries, 298–303, 300*f*
  computable general equilibrium (CGE), 298–302
  constant absolute risk aversion (CARA), 309, 312
  constant relative risk aversion (CRRA), 301–2
  definition, 299
  estimation, not direct elicitation, 326–27
  estimation procedures
    comparative statics, verification of, 415–16
    observations, 415–16
    structural estimation of preference parameters, 416
  randomization, 300, 333*n*26
  risk aversion, 303–16
  risk preferences, stability of, 324
polygraph test, 189–91
Porter, David, 33*n*5, 355*n*8
Post, T., 280
"postdictiveness," 63
Potamites, E., 361
Potters, Jan, 326, 376, 401
predictive and explanatory theory, distinctions, 62–67
  examples, 63–66
  k-level theory, 65–66
  making explanatory theories predictive, 66–67
  Nash *vs.* quantal response, 63–65, 64*t*
  "postdictiveness," 63
  unique mixed strategy equilibrium, 65
  zero-sum game, 65
predictive power of theories, comparing, 120–21
pricing, 281–82
  hedonic pricing method (HPM), 329
  option pricing by students and professional traders, 381–82, 382*t*
prisoner's dilemma, 15, 55, 152, 153*t,* 170, 210*t,* 211*t,* 217–18, 223
professional rules of conduct and subject surrogacy, lobbying experiment, 376–77, 376*t*
professionals *vs.* students, 8, 360–86
  behavioral investigation, 381–82, 382*t*
  businessmen as subjects in sealed offer markets, use of, 373–74, 373*t*
  common value offer auctions, comparison of naive and experienced bidders, 378–80
  comparative statics, 364, 371
  conclusions on, 384–86
  construct validity, 363
  coordination failure, experienced managers, 383–84, 383*t,* 384*t*
  ecological validity, 362, 364
  experientia docet, 380–81
  external validity, 362, 363–64
  information signals, 374–78
  internal validity, 362–63
  market experiments, 370–73
  methodological concepts defined, 362–64
  minimax model for soccer players, 380–81, 380*t*
  modeling caveats, 364–65
  observed differences, 385
  other regarding preferences, 365–70
  statistical conclusion validity, 363
  summary of the distance to the theoretical prediction, 384–85, 385*t*
  trust game, Costa Rica, CEO's compared to students, 220, 369, 443–44
  ultimatum game, 364–66, 387*n*11
  undergraduate students, recruiting, 361–62
  unique mixed strategy equilibrium, 380
  winner's curse, field experiments, 344–45
  zero-sum game, 380–81
promise of experimental economics, 2, 13–33
  Allais paradox, 14, 16
  bargaining behavior, 18–20
  building a research community, 14–15
  conclusions on, 32–33
  critiques and criticisms of experiments, 30–32
  ecological validity, 29, 31
  experimental economists and laboratories, statistics, 14
  external validity, 8–9
  fMRI (functional magnetic resonance imaging), 29
  heritability of economics, 30
  individual choice behavior, 16–18
  market design
    cybernomics, 33*n*5
    designing labor markets for doctors, 22–29
    first attempts, 21
    as ongoing process, 22
  Nash bargaining theory, 18–20
  new areas of research, 29–30
  perfect equilibrium, 14, 19–20
  prospect theory, 16–17
  speaking to theorists and searching for facts, 15
  utility theory, 16–18
  whispering in the ears of princes, 20–22
promise of generalizability, 252
  external validity of lab experiments, misleading emphasis on quantitative effects, 395–96
proofreading and exam grading, 275–76
propensity score matching, 423, 423*t*
pro-social behavior, 208, 213–25, 227*n*12
  anonymity, 284
  individual self-selection and, 264–66, 274–75
  lab–field generalizability in experimental economics, 283–85
  scrutiny and, 451, 456
  volunteering, 445, 458*n*14
pro-sociality, 283–85

prospect theory, 16–17, 152, 155, 162, 325
Pruitt, D., 255
psychology and economics, 29–30
  convergence and differences (*See* psychology and economics, convergence and difference)
  research (*See* psychology research and economics research)
  social psychology (*See* social psychology and economics)
psychology and economics, convergence and difference, 5, 6, 181–94, 200–204
  behavioral economics, 182–84, 185, 187–88
  context in experimental design, 202
  deception, 186–88, 202–3
    Good Samaritan parable, 203
  discussion of, 6, 200–204
  ecological validity, 187–88, 201
  importance of psychological studies, generally, 182–83
  monetary incentives, 203
  neuroeconomics, 188–94
    construct validity, 189–91, 193–94
    fMRI (functional magnetic resonance imaging), 188, 190–91, 192, 193
    hard science, 189–91
    neuroimaging and behavioral economics, 191–94
    polygraph test, 189–91
    "true" feelings inferred from physiology, 189–91
    trust, 182, 190–93
    value, 192–93
  observable behavior, 182–84
  participant bias, 185–88, 201–3
  primary goal of psychology, 183
  "psychologizing" and lack of supporting evidence, 184
  real world experiments, 185, 187
  ultimatum game, 186, 188
  unified approach, 194, 203–4
  unobtrusive measurement, 186–87, 201
psychology research and economics research
  behavioral economics, 152–54
  differences between, 151–52
  1-800 critique, 5, 151–63
  other regarding preferences, 152, 157
  parsimony and descriptive accuracy, trade-off between, 151–52, 163
  Pesendorfer's assertion, 153–54
publication trends, 133–36
public goods game, 211*t,* 217, 221, 283, 284, 317, 367
Puppe, Clemens, 222
"pure choice," 97–98
pure strategy experiments, results, 77–78, 77*f*–79*f*
qualitative external validity *vs.* quantitative results, 393–400, 402*n*4
qualitative results, external validity of, 443
quantal response equilibrium, 44, 50–53, 53*f,* 55, 62, 63–65, 64*t*
quantitative effects, misleading emphasis on. *See* external validity
quantitative predictions, 159–62

Rabin, Matthew, 54, 157, 226*n*4
Radner, R., 47
randomization, 225, 274, 303, 331
  bias, 303, 323–24, 427, 439
  by computers, 331*n*13
  control trials, origination, 333*n*26
  Denmark, policy lotteries, 300, 323–24, 333*n*26
  empirical methods, 422, 424, 426–27, 430, 438
  field experiments, 250, 300, 424
  participation decision, 438–39
randomized response technique, 216
Rapoport, Amnon, 66, 219, 255, 452
rare events, weighting, 152, 154, 155–56, 155*t,* 156*t,* 160, 162
Rassenti, S.J., 21
Rasul, Iwan, 215
realism, 63, 177–78, 253–54, 408, 414, 418
real world
  data, 63, 68
  decisions, 408
  economic theory and, 60, 68, 145
  experiments and, 145–47, 185, 187, 408–9, 411–12, 421–22, 432, 434
  lab experiments, measuring social preferences, 207–25, 249–86
  policy changes, 414–18
real world experiments and, 432, 434
reconciliation, testing theories in the lab, 70–72
Reiley, David, 9, 332*n*19, 333*n*25, 333*n*28, 334*n*31, 402*n*5, 414
replication, 5, 252, 256–57, 288*n*19, 422, 440–42, 441*f,* 448–49, 449, 452, 458*n*11
  external validity of lab experiments, misleading emphasis on quantitative effects, 402*n*2
  social psychology and economics, 171, 173, 175, 178
  testing theories in the lab, 71
reputation, consequences and concerns, 214, 218, 223, 270, 285, 349, 436, 448, 459*n*17
research community, building, 14–15
research papers, publication trends, 133–34, 133*t*
"reservation-based" search (RBS), defined, 87
Rey-Biel, Pedro, 4
Richerson, P.J., 272, 288*n*21
Riedl, Arno, 211
Riker, William H., 386*n*3
risk and time preference, 282
risk aversion, policy lotteries, 303–16

risk-neutral Nash equilibrium (RNNE), 116–17, 342, 344
Roberts, S., 72
Rockenbach, Bettina, 222, 381–82
role of experiments, 145–46
role of theory in experiments, 60–62
  expected utility theory, 61–62
  quantal response theory, 62
  ultimatum game, 61
Roll, R., 341
Rondeau, Daniel, 434, 450
Rose, S.L., 343
Rosen, S., 83*n*6
Rosenbaum, P.R., 423
Rosenthal, Robert W., 44, 220, 226*n*9
Rosnow, Ralph L., 220, 226*n*9
Ross, Lee, 215, 217
Ross, Thomas, 166–68
Roth, Alvin E., 1–2, 13, 16, 18–20, 22–28, 34*n*12, 46, 76–77, 117, 121, 128*nn*9–10, 152, 202, 211, 217, 219, 285, 402*n*5
Rubin, D.B., 423
Rubinstein, A., 19, 65, 102*n*5
Runkel, Philip, 175, 177–78, 198
Rupp, Nicholas G., 227*n*11
Rustagi, D., 284
Rutström, E.Elisabet, 7, 9, 305, 311, 317, 325, 334*n*29, 407

Sadiraj, V., 316
Sadrieh, Abdolkarim, 222
Salant, Y., 102*n*5
Samuelson, Larry, 100, 144, 218, 249, 341, 355*n*6
Sandholm, W.H., 54
satisficing model, 87, 91, 96–97
Schelling, Thomas, 33*n*2
Schipper, Burkhard, 34*n*11
Schmidt, K.M., 20, 49, 54, 61, 157
Schmittberger, Rolf, 211
Schnier, K., 280
Schotter, Andrew, 3, 60, 68, 71, 73–74, 77, 80–82, 83*n*1, 128*n*9
Schoumaker, F., 19
Schram, Arthur, 255, 284, 288*n*26, 395, 397–98
Schubert, R., 305
Schultz, Duane P., 213, 262
Schwarze, Bernd, 211
scientific view, 6, 251, 252–54, 395
  external validity, origination of term, 253–54
  game show field data, 254
  parallelism, 252–53, 263, 396, 453, 456
  promise of generalizability, 252
  realism, 253–54
scrutiny, 213–15, 224
sealed-envelope-submission strategy, journal publication, 143
Sears, D.O., 186
Seasholes, M., 288*n*23
second-price sealed bid auctions, 110, 113, 119
Seki, Erika, 283, 369, 385
Sela, A., 73
selection bias, 266, 426–27, 438, 443, 445–47
self-confirming equilibrium, 48–49
selfish preferences theory, 44, 46–47, 52–53
self-selection of the individuals making decisions, 264–66
Selten, Reinhard, 14, 33*n*2, 54
Serneels, Pieter, 281
Serra, D., 283–84
Shachat, Keith, 216, 226*n*6
Shadish, William R., 387*n*8
Shaffer, V.A., 261
Shang, Jen Y., 401
Shapley, Harlow, 23, 54, 208
Shogren, Jason F., 219
short-run decisions, 222, 329
short-run views, 421
signals, bid classification, 111–12, 111*f*, 113*f*
silent auctions, 280
Silverthorne, Colin, 219
Simon, Herbert, 86, 91, 99
simplicity of experiments, 147
Skinner, B.F., 167, 182
Slonim, Robert, 219, 265, 286, 445–47
Smith, Adam, 209
Smith, Vernon L., 14, 21, 33*n*5, 45, 147, 216–17, 226*n*3, 226*n*6, 252, 255, 266, 285, 327, 387*n*4, 433, 458*n*7, 459*n*18
Snowberg, E.C., 279
soccer, 274, 386, 410, 414, 452
  professionals *vs.* students, 380–81, 380*t*
social conflict experience, 156–59, 157*t*, 158*f*, 159*t*
socially efficient or free-riding equilibrium, 367–68, 368*f*
social preferences, measuring. *See* measuring social preferences, lab experiments
social psychology and economics, 5, 166–80
  anchoring-and-insufficient-adjustment bias, 171–73
  anchoring effects, reversal of, 172–73
  auction pricing, 173
  comparison framework, major research strategies, 175, 176*f*
  conclusions on, 178–79
  deception, 168–69
  decision bias, 171–72
  definitions, 167
  economics approach, advantages, 169–70
  external validity, 171, 175–76, 178, 197
  field experiments, 197–98
  general model for experimental inquiry, 5, 166–79
  goals, 167–68
  hammer and the screwdriver, 6, 197–98

social psychology and economics, (*continued*)
historical perspectives, 167–68
hot hand in basketball, 171–72
laboratory and field experiments
best uses view, 197–98
as complimentary, 175–77
methods and approaches, 168–69
monetary incentives, 168–69
multiple methods, 174–78, 197–98
obtrusive-unobtrusive research operations, 175, 176*f,* 198
replication, 171, 173, 175, 178
research strategies described, 177*t*
social psychology's advantages, 170–74
study of emotions, 174
universal behavior, 175, 176*f,* 198
Soetevent, Adriaan R., 212, 401
Solow, R.M., 352
Sonneman, U., 278–79
Sonnemans, J., 202
Sopher, B., 83*n*1
Sotomayor, M., 22
speaking to theorists and searching for facts, 15
sports card trading, 214, 220, 224, 251, 255–56, 264–65, 268–72, 269*t,* 270*f,* 271*t,* 346–47, 397, 401*n*2, 451, 459*n*17
sports good and consumer good trading, 277–78
Sprenger, C., 282
stable matchings theory, 22–23
stag hunt game, 157–59, 157*t,* 158*f*
Stahl, D.O., 54, 65
stakes, 219, 266–67
stakes, morality and wealth components, 219
standard oral double auction (ODA), 370
Stanley, Julian, 176, 187, 253, 272, 285
stated choice method (SCM), 329
static *vs.* comparative approach, 68–69
statistical conclusion validity, 363
statistical specification, 310–11, 356*n*26
Staw, B., 171
Stein, William E., 219
Stern, N., 301
Sternberg, S., 151
sticky downward wages, 352–53
Stoop, Jan, 252, 274, 276–77, 451
Strazzera, E., 332*n*20
structural approach, testing theories in the lab, 67–68
student funds, donations to, 273–74
students
undergraduate students, recruiting, 361–62
students v. professionals. *See* professionals *vs.* students
study participants, selection, 219–21
subgame perfect equilibrium, 46, 83*n*2, 157, 370
subject classification, 111–12, 111*t,* 113*t*
subject pool bias, 220
subject pools. *See* professionals *vs.* students
summary of experimental games, 210*t*–211*t*
Sunder, Shyam, 370, 386*n*2
Swedish lotteries, 279

Tanaka, T., 284
Taylor, Laura O., 215, 226*n*6
Tenorio, R., 282
testing for selection, 115–16
testing theories in the lab, 67–76
comparative *vs.* static approach, 68–69
data features and goodness of fit, 77–82
dropouts, 74–75
effort level, 74
external validity, 71
formalization of the difference, 69–70
goodness-of-fit measure, 72–73, 77–82
individual's choices, 77, 79*t*
lessons to be learned from experiments, 72–73
level of aggregation, 73–76, 75*f,* 76*f*
pure strategy experiments, results, 77–78, 77*f*–79*f*
reconciliation, 70–72
replication, 71
structural approach, 67–68
Thaler, Richard H., 155, 211, 259, 287*n*7, 444
theoretical models, 146–47
theorists and experimentalists, generally, 132–33
theorists dabbling in experiments, 138, 138*f,* 139*n*8
Thibaut, J., 170, 178
Thurstone, L.L., 16
Todd, P.M., 327
"toothbrush theories," 120
Tougareva, E., 219
tournaments and contests
competition and gender, tournament selection example, 122–26, 125*f,* 283, 440
prize structure, 73–76
travel cost method (TCM), 329
treatment-driven experiments, 121–26
channel driven results, 121
competition and gender, tournament selection example, 122–26, 125*f,* 283, 440
Trivers, R.L., 146
"true" feelings inferred from physiology, 189–91
trust, 182, 190–93
trust game, 184, 210*t,* 283–84
Costa Rica, CEO's compared to students, 220, 369, 443–44
discount rates, 317
generally, 210*t*
subject pool bias, 220
volunteer bias, 265
Tufano, F., 440
Tversky, Amos, 16–17, 29, 55, 61, 65, 155–56, 161, 171–72, 327
two-player games, 108, 119

common value auction, 110
second-price sealed bid auctions, 110, 113, 119
unique mixed strategy equilibrium, 121
zero-sum games, 121
Tyler, Tom R., 5, 185–86, 188, 200–201, 203

Udry, Chris, 287*n*4
ultimatum game, 46–49, 47*t,* 61, 83*n*3, 152, 153*t,* 157, 186, 188, 201, 210*t,* 217–19, 288*n*18, 317, 364–66, 387*n*11
undergraduate students, recruiting, 361–62
unifying theories approach, 146
unique mixed strategy equilibrium, 65, 121, 380, 451
universal behavior, 175, 176*f,* 198
unobtrusive measurement, 186–87, 201
unobtrusive-obtrusive research operations, 175, 176*f,* 198
utility maximization model, 209, 211–12
utility theory, 16–18

value, 192–93
Van Huyck, John, 384
van Winden, Frans, 376
Varian, H.L., 147
Verhoogen, Eric, 219
Vesterlund, Lise, 8, 17, 30, 122, 126, 221, 227*n*10, 279, 400, 415, 440, 443–44
Vinod, H.D., 300
virtual experiments, 327–29
Volij, Oscar, 274, 380, 385, 387*n*12, 410, 452
voluntary contribution mechanism (VCM), 274–75, 284, 369–70, 451
volunteer bias, 265
von Neumann, J., 61, 152, 175
voting experiment, 44–45, 45*f,* 50–52, 51*f*

wages. *See* gift exchange, experimental labor markets
Walker, James M., 227*n*9, 266, 459*n*18
Wallis, Allen, 16, 34*n*12
Ward, Andrew, 217
Wason, P.C., 410
"The Weakest Link" (television game show), 223, 280–81
Webb, E.J., 186
Weber, R.J., 343
Weber, Roberto A., 221, 287*n*8
Weigelt, Keith, 211
Weisberg, D., 190
whispering in the ears of princes, 20–22
Wilcox, N.T., 313–14
Wilson, P.W., 54, 65
Wilson, Robert, 21
Winking, Jeffrey, 449
winner's curse, 7–8, 340–47, 409–12
adverse selection effect, 340, 342–43, 346–47, 354*n*3
common-value auctions, 110, 112–13, 115
conclusions on, 353–54
construction industry bidding, 409–10
ecological validity, 344
external validity, 340, 343, 344–47
field experiments, 344–47
initial results, relation to theory and experimental methodology, 340–44
Outer Continental Shelf (OCS) oil lease auctions, 340, 343, 345, 347, 354*n*3
overbidding, 342–43
risk-neutral Nash equilibrium (RNNE), 342, 344
Wolfers, J., 279
Wooders, J., 387*n*12
wool buyers and students, simulated progressive auction, 371–73
world outside the lab, 396–99

Yafe, H., 215
Yariv, Leeat, 4, 137
Yellen, J., 352
Yoeli, E., 450

Zeitlin, A., 283
zero-sum game, 65, 121, 380–81, 451
Zhang, B., 361
Zheng, Jie, 3, 83*n*1
Zhong, C., 170
Zwick, Rami, 226*n*6